Department of Statistics, Universidad Carlos III de Madrid

PhD Thesis

Flexible Bayesian Nonparametric Priors and Bayesian Computational Methods

Weixuan Zhu

Advisors:

Fabrizio Leisen
University of Kent

Juan Miguel Marín
Universidad Carlos III de Madrid

Contents

List of Figures

List of Tables

Abstract

The definition of vectors of dependent random probability measures is a topic of interest in Bayesian nonparametrics. They represent dependent nonparametric prior distributions that are useful for modelling observables for which specific covariate values are known. Our first contribution is the introduction of novel multivariate vectors of two-parameter Poisson-Dirichlet process. The dependence is induced by applying a Lévy copula to the marginal Lévy intensities. Our attention particularly focuses on the derivation of the Laplace functional transform and the analytical expression of the Exchangeable Partition Probability function (EPPF). Their knowledge allows us to gain some insight on the dependence structure of the priors defined. The second part of the thesis deals with the definition of Bayesian nonparametric priors through the class of species sampling models. In particular, we focus on the novel Beta-GOS model introduced by Airoldi, Costa, et al. (2014). Our second contribution is the modification of the Beta-GOS model with the motivation to accommodate both temporal and spatial correlations that exist in many applications. We then apply the modified model to simulated fMRI data and display the results. Finally, we aim to give contribution to another popular area of nonparametric computational methods in Bayesian inference: Approximate Bayesian Computations (ABC), by providing a new sampler BC_{bl}. It combines the idea of standard ABC and bootstrap likelihood and allows to avoid the choice of ABC parameters. Our work is actually inspired by a recent algorithm BC_{el} proposed by Mengersen, Pudlo and Robert (2013) that uses the well-established empirical likelihood approximation. However, to ensure that the empirical likelihood converges to the true likelihood, it requires a very careful choice of the constraints. This choice is not clear in many cases. On the other hand, the bootstrap likelihood is an automatic procedure, with only a few trivial parameters to specify. The advantages of our algorithm BC_{bl} are illustrated with several examples.

Chapter 1

Introduction

Consider an infinite sequence of observations $X^{(\infty)} = (X_n)_{n\geq 1}$, defined on the probability space $(\Omega, \mathscr{F}, \mathbb{P})$ and taking values on a measurable space $(\mathbb{X}, \mathcal{X})$, with $\mathbb{X}$ a Polish Space and $\mathcal{X}$ the Borel σ–algebra of subsets of $\mathbb{X}$. Furthermore, denote by $\mathcal{P}_{\mathbb{X}}$ the space of all probability measures supported on $\mathbb{X}$.
One way of justifying Bayesian approaches to inference is through exchangeability and de Finetti's (1937) representation theorem. A sequence of random variable $(X_n)_{n\geq 1}$ is said to be exchangeable if for any $n \geq 1$ and permutation of σ of $\{1, \ldots, n\}$ we have

$$(X_1, \ldots, X_n) \stackrel{d}{=} (X_{\sigma(1)}, \ldots, X_{\sigma(n)})$$

so that the order of the sampling scheme does not affect the distribution. A sequence of independent and identically distributed variables is exchangeable but the concept of exchangeability is more general. For example, sequences sampled without replacement are also exchangeable although the sampled variables are not independent. De Finetti's theorem states that a sequence of random variables is exchangeable if and only if it is a mixture of sequences of independent and identically distributed (i.i.d.) random variables.

Theorem 1 *(De Finetti, 1937) The sequence $X^{(\infty)}$ is exchangeable if and only if there exists a probability measure Q on $\mathcal{P}_{\mathbb{X}}$ such that, for any $n \geq 1$ and $A = A_1 \times A_2 \times \cdots \times A_n \times \mathbb{X}^{(\infty)}$,*

$$P[X^{(\infty)} \in A] = \int_{\mathcal{P}_{\mathbb{X}}} \prod_{i=1}^{n} p(A_i) Q(dp)$$

where $A_i \in \mathcal{X}$ for any $i = 1, \ldots, n$ and $\mathbb{X}^{(\infty)} = \mathbb{X} \times \mathbb{X} \times \cdots$

De Finetti's theorem also implies that a sequence of random variables is exchangeable if and only if they are conditionally independent and identically distributed (i.i.d.), and that the unknown , i.e. $\tilde{p}$ (that allows us to see such a sequence as an i.i.d sample) should be random and measured with a distribution Q, namely the prior distribution. In fact, it can be proved that the measure Q, sometimes called the *de Finetti measure*, is uniquely determined for a given exchangeable sequence.
In the Bayesian literature, and hereafter in this thesis, we will represent this kind of model through the conditional dependence structure,

$$\begin{aligned} X_i|\tilde{p} &\overset{iid}{\sim} \tilde{p} \quad i = 1, 2, ..., n \\ \tilde{p} &\sim Q \end{aligned} \tag{1.1}$$

Whenever Q is degenerate on a subset of $\mathcal{P}_{\mathbb{X}}$ indexed by a finite dimensional parameter, we say that inference is *parametric*. On the other hand, when no restriction is made, or there is a restriction to infinite-dimensional subspaces of $\mathcal{P}_{\mathbb{X}}$, the model is then *nonparametric*. This work will focus on nonparametric problems, and we start by giving the definition of two famous examples: Dirichlet process (DP) and two-parameter Poisson Dirichlet process (PD).
Ferguson (1973) introduces the idea of Dirichlet process - a probability distribution on the space of probability measures. We start by reviewing the definition of Dirichlet distribution.

Definition 1 *We say that a continuous, $K-$variate random variable $\mathbf{X} = (X_1, \ldots, X_K)$ follows a Dirichlet distribution with parameters $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_K)$ where $\alpha_i > 0$ for $i = 1, \ldots, K$ if it has a probability density function given by:*

$$f(\mathbf{x}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{i=1}^{K} \alpha_i)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \prod_{i=1}^{K} x_i^{\alpha_i - 1} \tag{2.1}$$

for $0 < x_i < 1$ for $i = 1, \ldots, K$ and $\sum_{i=1}^{K} x_i = 1$. In this case, we write $\mathbf{X} \sim Dir(\alpha_1, ..., \alpha_K)$.

Notice that when $K = 2$, the Dirichlet distribution reduces to a beta distribution. Given the above definition, we can now define the Dirichlet process.

Definition 2 *Let $\mathbb{X}$ be a set and let $\mathcal{X}$ be a σ-field of subsets of $\mathbb{X}$. Let $\tilde{\alpha}$ be a non-null finite measure (non-negative and finitely additive) on $(\mathbb{X}, \mathcal{X})$. We say P*

is a Dirichlet process on $(\mathbb{X}, \mathcal{X})$ *with parameter* $\tilde{\alpha}$ *if for every* $k = 1, 2, ...,$ *and measurable partition* $(B_1, ..., B_k)$ *of* $\mathbb{X}$*, the distribution of* $(P(B_1), ..., P(B_k))$ *is a Dirichlet distribution* $Dir(\tilde{\alpha}(B_1), ..., \tilde{\alpha}(B_k))$.

In the literature, it is common to specify the DP through two parameters, that is $\alpha = \tilde{\alpha}(\mathbb{X})$, the total mass and $G_0 = \frac{\tilde{\alpha}(\cdot)}{\tilde{\alpha}(\mathbb{X})}$, the base measure. We denote it as $\mathrm{DP}(\alpha, G_0)$ in this thesis.

There are several equivalent definitions of the Dirichlet process. Besides the definition above, the Dirichlet process can be represented by a Polya urn scheme, and it shows that draws from the DP are both discrete and exhibit a clustering property. Thus, suppose that $X|P \sim P$ and $P \sim \mathrm{DP}(\alpha, G_0)$ as previously. One can obtain a representation of the distribution of the X_{n+1} in terms of successive conditional distributions of the following form:

$$X_{n+1} \mid X_1, ..., X_n \ \sim \ \frac{1}{n+\alpha} \sum_{j=1}^{n} \delta_{X_j} + \frac{\alpha}{n+\alpha} G_0$$

Here, δ_x is the distribution concentrated at the single point x. So the distribution of X_{n+1} can be described as X_{n+1} being one of the previous X_j's with probability $\frac{1}{n+\alpha}$ and getting a new draw from G_0 with probability $\frac{\alpha}{n+\alpha}$. Taking into account that many of the previous X_j's are equal among themselves, the conditional draw can be characterized as setting to θ_j with probability $\frac{n_j}{n+\alpha}$, where the θ_j are distinct values of $\{X_1, ..., X_n\}$ with frequencies n_j respectively, $j = 1, ..., k$, and a new draw from G_0 with probability $\frac{\alpha}{n+\alpha}$:

$$X_{n+1} \mid X_1, ..., X_n \ \sim \ \begin{cases} \delta_{\theta_j}, & \text{with probability } \dfrac{n_j}{n+\alpha} \quad \text{for } j = 1, ..., k \\ G_0, & \text{with probability } \dfrac{\alpha}{n+\alpha} \end{cases} \tag{1.2}$$

where k is the number of distinct observations in $\{X_1, ..., X_n\}$.

This representation is important to address posterior inference with a Gibbs sampling algorithm when DP is assumed.

Many generalizations of the Dirichlet process have been proposed, and among them, an extension called two-parameter Poisson-Dirichlet process (also known as Pitman-Yor process) with parameters (σ, θ), introduced in Pitman (1995), has gained more and more popularity. It is motivated by the limitation of DP that the probability weights defined in (1.2) do not depend on the number of clusters in which the data are grouped.

The two-parameter Poisson Dirichlet process, which we denote as $\text{PD}(\sigma, \theta)$, can be characterized by the predictive conditional distributions, similarly as the Dirichlet process. And the predictive distribution of $\text{PD}(\sigma, \theta)$ is:

$$X_{n+1} \mid X_1, ..., X_n \sim \begin{cases} \delta_{\theta_j}, & \text{with probability } \dfrac{n_j - \sigma}{n + \theta} \quad \text{for } j = 1, ..., k \\ G_0, & \text{with probability } \dfrac{\theta + \sigma k}{n + \theta} \end{cases} \tag{1.3}$$

where k is the number of distinct observations in $\{X_1, ..., X_n\}$ and the θ_j's are distinct values of $\{X_1, ..., X_n\}$ with frequencies n_j respectively, $j = 1, ..., k$. Hence, the probability of obtaining new values is monotonically increasing in k and the value of σ can be used to tune the strength of the dependence on k. The $\text{PD}(\sigma, \theta)$ process yields a more flexible model for clustering than the one provided by the Dirichlet process. Indeed when $\sigma = 0$, the $\text{PD}(\sigma, \theta)$ process reduces to a Dirichlet process. There is a growing literature about nonparametric priors that could be expressed with predictive distribution similar with (1.2) and (1.3), known under the name species sampling sequence (SSS). Details will be provided later.

Another popular construction of $\text{DP}(\alpha, G_0)$, namely stick-breaking construction, involves representing the random probability measure $\tilde{p}$ as:

$$\tilde{p} = \sum_{i=1}^{\infty} W_i \delta_{X_i} \tag{1.4}$$

where the weights W_i and the atoms X_i are random and independent for all i. The values of X_i are independent draws from the base measure G_0 and the weights W_i are defined as:

$$W_1 = V_1, \qquad W_i = V_i \prod_{j=1}^{j-1} (1 - V_j), \ j \geq 2$$

where $V_i \sim \text{Beta}(1, \alpha)$. The name "stick-breaking" comes from the definition of the weights W_i, which can be though as the length of the piece of a unit-length stick assigned to the i-th value. The two-parameter Poisson Dirichlet process can also be constructed with the same procedure. In fact, Pitman (1995) proves that if $V_i \sim \text{Beta}(1 - \sigma, \theta + i\sigma)$, where $\sigma \in (0, 1)$ and $\theta > -\sigma$, then the random probability measure $\tilde{p}$ is a $\text{PD}(\sigma, \theta)$. From this representation, we can also see that DP is a special case of $\text{PD}(\sigma, \theta)$ by setting $\sigma = 0$.

This representation is very attractive from a computational point of view, mainly

because it suggests an intuitive way to generate realization from $\tilde{p}$ with an appropriate truncation of (1.4). However, the analytic expressions for some quantities of interest are not available if we resort to the stick-breaking representation. For instance, the exchangeable partition probability function (EPPF) which gives information about the clustering behavior of the prior and allows to compute the predictive distributions. The use of completely random measures in Bayesian nonparametrics provides a tool that is useful both for the understanding of the behavior of commonly exploited priors and for the development of new models. So it is definitely worth to explore the potential of random probability measures in BNP since it allows us to find many alternatives to the Dirichlet process which still maintain mathematical tractability. In fact, many of the priors in Bayesian nonparametrics can be constructed based on suitable transformations of completely random measures. For example, the Dirichlet process itself can be seen as the normalization of the so-called gamma completely random measure.

1.1 Completely random measure

We first review the definition of completely random measure give by Kingman (1967). Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(\mathbb{X}, \mathcal{X})$ a measurable space, with $\mathbb{X}$ a Polish Space and $\mathcal{X}$ the Borel σ–algebra of subsets of $\mathbb{X}$. Let $\boldsymbol{M}_{\mathbb{X}}$ be the space of all boundedly finite measures on $(\mathbb{X}, \mathcal{X})$ endowed with the Borel σ–algebra $\mathcal{M}_{\mathbb{X}}$.

Definition 3 *Let $\tilde{\mu}$ be a measurable mapping from $(\Omega, \mathcal{F}, \mathbb{P})$ into $(\boldsymbol{M}_{\mathbb{X}}, \mathcal{M}_{\mathbb{X}})$ and such that for any $A_1, \ldots, A_n$ in $\mathcal{X}$, with $A_i \cap A_j = \emptyset$ for any $i \neq j$, the random variables $\tilde{\mu}(A_1), \ldots, \tilde{\mu}(A_n)$ are mutually independent. Then $\tilde{\mu}$ is called a completely random measure (CRM).*

A CRM on $\mathbb{X}$ can always be represented as the sum of two components: a completely random measure $\tilde{\mu}_c = \sum_{i=1}^{\infty} J_i \delta_{X_i}$, where both the positive jumps J_i and the $\mathbb{X}$-valued locations X_i are random, and a measure with random masses at fixed locations. Specifically,

$$\tilde{\mu} = \tilde{\mu}_c + \sum_{i=1}^{M} V_i \delta_{x_i}$$

where the fixed jump points $x_1, ..., x_M$ are in $\mathbb{X}$, the nonnegative random jumps $V_1, ..., V_M$ are mutually independent and they are independent from $\tilde{\mu}_c$. Finally, $\tilde{\mu}_c$

is characterized by the Lévy-Khintchine representation, which states

$$\mathbb{E}\left[e^{-\int_{\mathbb{X}} f(x)\tilde{\mu}_c(dx)}\right] = \exp\left\{-\int_{\mathbb{R}^+\times\mathbb{X}} [1-e^{-yf(x)}]\bar{\nu}(dy,dx)\right\}$$

where $f : \mathbb{X} \to \mathbb{R}^+$ is a measurable function such that $\int |f| d\tilde{\mu}_c < \infty$ (almost surely) and $\bar{\nu}$ is a measure on $\mathbb{R}^+ \times \mathbb{X}$ such that $\int_B \int_{\mathbb{R}^+} \min\{y, 1\}\bar{\nu}(dy, dx) < \infty$ for any B in $\mathcal{X}$. The measure $\bar{\nu}$ characterizing $\tilde{\mu}_c$ is referred to as the *Lévy intensity* of $\tilde{\mu}_c$. It contains all the information about the distribution of jumps and locations of $\tilde{\mu}_c$. It is useful to separate the jump and location part of $\bar{\nu}$ by writing it as

$$\bar{\nu}(dy, dx) = \nu_x(dy)\alpha(dx)$$

where α is a measure on $(\mathbb{X}, \mathcal{X})$ and ν is a measure on $\mathbb{R}^+$ such that

$$\int_{\mathbb{R}^+} \min(y, 1)\, \nu(dy) < \infty.$$

For our purpose, we will focus on the homogeneous case, i.e. $\nu_x = \nu$ for any x, where the distribution of the jumps of $\tilde{\mu}_c$ is independent of the location. The following are two famous examples of CRM.

Example 1 *Let α be a finite non null measure on $(\mathbb{X}, \mathcal{X})$. A CRM $\tilde{\gamma}$ whose Lévy intensity is given by*

$$\nu(dy, dx) = \frac{e^{-y}}{y} dy\alpha(dx) \tag{1.5}$$

is a gamma process with parameter measure α on $\mathbb{X}$. If we set the measurable function $f = \lambda \mathbb{1}_A$ with $\lambda > 0$, $A \in \mathcal{X}$ and $\mathbb{1}$ the indicator function, it is characterized by its Laplace functional which is given by

$$\mathbb{E}[e^{-\lambda\tilde{\gamma}(A)}] = [1+\lambda]^{-\alpha(A)}$$

.

Example 2 *Let α be a finite non null measure on $(\mathbb{X}, \mathcal{X})$ and $\sigma \in (0, 1)$. Consider a CRM $\tilde{\mu}_\sigma$ with Lévy intensity*

$$\nu(dy, dx) = \frac{\sigma}{\Gamma(1-\sigma)} y^{-1-\sigma} dy\alpha(dx) \tag{1.6}$$

Then $\tilde{\mu}_\sigma$ is a σ-stable process with parameter measure α on $\mathbb{X}$. Similarly, the Laplace transform of $\tilde{\mu}_\sigma(A)$ is $\mathbb{E}[e^{-\lambda\tilde{\mu}_\sigma(A)}] = e^{-\lambda^\sigma\alpha(A)}$

In fact, as we have mentioned before, a random probability measure can be defined by a CRM. We firstly review the definition of normalized random measures in the following,

Definition 4 *(Normalized random measures) Let $\tilde{\mu}$ be a completely random measure on $(\mathbb{X}, \mathcal{X})$ such that $\tilde{\mu} < \infty$. We call the random probability measure $\tilde{p}$ a normalized random measure if $\tilde{p} = \frac{\tilde{\mu}}{\tilde{\mu}(\mathbb{X})}$.*

It is worth to mention that there is another construction of the Dirichlet process which involves normalizing a gamma process with intensity measure α. Kingman (1975) also studied the normalized σ-stable process by normalizing a σ-stable CRM. See the following two examples.

Example 3 *Let $\tilde{\gamma}$ be a gamma CRM with Lévy intensity* (1.5)*, with $0 < \alpha(\mathbb{X}) < \infty$. The random probability measure $p = \tilde{\gamma}/\tilde{\gamma}(\mathbb{X})$ is a Dirichlet process on $\mathbb{X}$ with parameter α.*

Example 4 *Let $\sigma \in (0, 1)$ and consider a CRM $\tilde{\mu}_\sigma$ with Lévy intensity* (1.6)*, with $0 < \alpha(\mathbb{X}) < \infty$. The random probability measure $p_\sigma = \tilde{\mu}_\sigma/\tilde{\mu}_\sigma(\mathbb{X})$ is a normalized σ-stable process with parameter σ and parameter measure α.*

1.2 Vectors of normalized random measures

The use of dependent CRMs in the construction of dependent random probability measures has been considered in many works. For sake of illustration, it helps to review some preliminary theories of vectors of CRM, see Leisen, Lijoi and Spano (2013) and Lijoi, Nipoti and Prünster (2014a).
Suppose $\tilde{\mu}_1, \ldots, \tilde{\mu}_d$ are completely random measures on $(\mathbb{X}, \mathcal{X})$ with respective marginal Lévy measures,

$$\bar{\nu}_i(\mathrm{d}x, \mathrm{d}y) = \alpha(\mathrm{d}x)\,\nu_i(\mathrm{d}y) \qquad i = 1, \ldots, d \tag{1.7}$$

The probability measure α on $\mathbb{X}$ is non–atomic (i.e. $\alpha(\{x\}) = 0$ for all $x \in \mathbb{X}$) and ν_i is a measure on $\mathbb{R}^+$ such that $\int_{\mathbb{R}^+} \min(y, 1)\,\nu_i(dy) < \infty$. Moreover, $\tilde{\mu}_1, \ldots, \tilde{\mu}_d$ are dependent and the random vector $(\tilde{\mu}_1, \ldots, \tilde{\mu}_d)$ has independent increments, in the sense that for any A_1,...,A_n in $\mathcal{X}$, with $A_i \cap A_j = \emptyset$ for any $i \neq j$, the random

vectors $(\tilde{\mu}_1(A_i), ..., \tilde{\mu}_d(A_i))$ and $(\tilde{\mu}_1(A_j), ..., \tilde{\mu}_d(A_j))$ are independent. This implies that for any set of measurable functions $\boldsymbol{f} = (f_1, \ldots, f_d)$ such that $f_i : \mathbb{X} \to \mathbb{R}^+$, $i = 1, \ldots, d$ are non-negative and $\int |f_i| \, d\tilde{\mu}_j < \infty$, one has a multivariate analogous of the Lévy-Khintchine representation of $(\tilde{\mu}_1, \ldots, \tilde{\mu}_d)$

$$\mathbb{E}\left[e^{-\tilde{\mu}_1(f_1) - \cdots - \tilde{\mu}_d(f_d)}\right] = \exp\left\{-\psi^*_{\rho,d}(\boldsymbol{f})\right\}. \tag{1.8}$$

where $\tilde{\mu}_i(f_i) = \int f_i d\tilde{\mu}_i$ and

$$\psi^*_{\rho,d}(\boldsymbol{f}) = \int_{\mathbb{X}} \int_{(0,\infty)^d} \left[1 - \mathrm{e}^{-y_1 f_1(x) - \cdots - y_d f_d(x)}\right] \, \rho_d(\mathrm{d}y_1, \ldots, \mathrm{d}y_d) \, \alpha(\mathrm{d}x) \tag{1.9}$$

and

$$\int_{(0,\infty)^{d-1}} \rho_d(\mathrm{d}y_1, \ldots, dy_{j-1}, A, dy_{j+1}, \ldots, dy_d) = \int_A \nu_j(\mathrm{d}y)$$

and ρ_d is the multivariate Lévy intensity. The form of the marginal Lévy intensity displayed in equation (1.7), entails that the jump heights of $(\tilde{\mu}_1, \ldots, \tilde{\mu}_d)$ are independent from the locations where the jumps occur. Moreover, these jump locations are common to all the CRMs and are governed by α. It is worth noting that, since $(\tilde{\mu}_1, \ldots, \tilde{\mu}_d)$ has independent increments, its distribution is characterized by a choice of $f_1, ..., f_d$, such that $f_i = \lambda_i \mathbb{1}_A$ for any set A in $\mathcal{X}$, $\lambda_i > 0$ for $j = 1, \cdots, d$, then

$$\psi^*_{\rho,d}(\boldsymbol{f}) = \alpha(A)\psi_{\rho,d}(\boldsymbol{\lambda}) \tag{1.10}$$

where $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_d)$ and

$$\psi_{\rho,d}(\boldsymbol{\lambda}) = \int_{(0,\infty)^d} \left[1 - \mathrm{e}^{-y_1\lambda_1 - \cdots - y_d\lambda_d}\right] \, \rho_d(\mathrm{d}y_1, \ldots, \mathrm{d}y_d)$$

We extend the definition of normalized random measures to the vectors of normalized random measures.

Definition 5 *Let $\tilde{\mu}_1, \ldots, \tilde{\mu}_d$ be a vector of CRMs on $\mathbb{X}$, and let $\tilde{p}_i = \frac{\tilde{\mu}_i}{\tilde{\mu}_i(\mathbb{X})}$, $i = 1, ..., d$, then the vector $(\tilde{p}_1, ..., \tilde{p}_d)$ is called a vector of dependent normalized random measures with independent increments on $(\mathbb{X}, \mathcal{X})$.*

1.2.1 Dependence through Lévy copulas

There are numbers of ways to introduce dependence between random measures $\tilde{p}_i$ and $\tilde{p}_j$ among the vector $(\tilde{p}_1, \cdots, \tilde{p}_d)$. One way to construct dependent random probability measures is to make use of dependent CRMs, which allows us to track some posterior properties analytically. From the definition, one can easily tell that the dependence between two random probability measures, say $\tilde{p}_i$ and $\tilde{p}_{i'}$, is induced by the dependence of the corresponding CRMs $\tilde{\mu}_i$ and $\tilde{\mu}_{i'}$. The use of dependent CRMs in the construction of dependent random probability measures has been considered in many works. Dependence could be induced through Lévy copulas at the level of Lévy intensity that allow to define multivariate Lévy intensities with fixed marginals, thus operating in a similar fashion as traditional copulas do for probability distributions. For example, in Leisen, Lijoi and Spano (2013), a vector of Dirichlet process is introduced by normalization of a vector of dependent gamma CRMs, whereas in Leisen and Lijoi (2011), a bivariate vector of two-parameter Poisson-Dirichlet processes is defined by suitable transformation of dependent stable CRMs. These two papers are in the same spirit, in terms of introducing dependence using Lévy copula, with the main difference being that the latter relies on the normalization of a random measure which is not completely random. Recently, Griffin and Leisen (2014) propose the Compound random measures which provides a unifying framework for previously proposed constructions of dependent random measures.
In this thesis, we extend the bivariate vector in Leisen and Lijoi (2011) to a more general case in a similar fashion. In the next section, we will see how to use Lévy copula to construct a multivariate vector of two parameter Poisson-Dirichlet process and prove some of its properties. For our purpose, it helps to review the definition of 2-dimensional Lévy copulas.

Definition 6 *A Lévy copula is a function $C : [0, \infty]^2 \to [0, \infty]$ such that*

1. *$C(y_1, 0) = C(0, y_2) = 0$ for any positive y_1 and y_2,*
2. *C has uniform margins, i.e. $C(y_1, \infty) = y_1$ and $C(\infty, y_2) = y_2$,*
3. *for all $y_1 < z_1$ and $y_2 < z_2$, $C(y_1, y_2) + C(z_1, z_2) - C(y_1, z_2) - C(y_2, z_1) \geq 0$.*

The definition in higher dimension is analogous (see Cont and Tankov (2004)). Let $U_i(y) := \int_y^\infty \nu_i(s)\,\mathrm{d}s$ be the i–th marginal tail integral associated with ν_i. If both the copula C and the marginal tail integrals are sufficiently smooth, then the

multivariate Lévy intensity ρ_d can be obtained through the following expression:

$$\rho_d(y_1,\dots,y_d) = \frac{\partial^d C(s_1,\dots,s_d)}{\partial s_1 \cdots \partial s_d}\bigg|_{s_1=U_1(y_1),\dots s_d=U_d(y_d)} \nu_1(y_1)\dots\nu_d(y_d). \tag{1.11}$$

A wide range of dependence structures can be induced through Lévy copulas. For example, the Lévy-Clayton Copula is defined by

$$C_\gamma(s_1,\dots,s_d) = (s_1^{-\gamma}+\cdots+s_d^{-\gamma})^{-\frac{1}{\gamma}} \qquad \gamma>0. \tag{1.12}$$

where the parameter γ regulates the degree of dependence. The bivariate Lévy intensity used in Leisen and Lijoi (2011) can be recovered by using this Copula with $\gamma=\frac{1}{\sigma}$. To be more specific, we illustrate the following examples.

Example 5 *Suppose $\tilde{\mu}_1$ and $\tilde{\mu}_2$ are σ-stable CRMs, i.e.*

$$\nu_i(dy) = \frac{\sigma}{\Gamma(1-\sigma)} y^{-1-\sigma}dy, \quad i=1,2$$

where $\sigma\in(0,1)$. Under the choice of Lévy-Clayton in (1.12) *with $d=2$ and $\gamma=\frac{1}{\sigma}$, one can obtain the bivariate Lévy intensity*

$$\rho_2(y_1,y_2) = \frac{\sigma(1+\sigma)}{\Gamma(1-\sigma)}(y_1+y_2)^{-\sigma-2}\mathbb{1}_{(0,+\infty)^2}(y_1,y_2)$$

by applying (1.11).

The following example shows the bivariate case of the multivariate vector of Dirichlet processes introduced in Leisen, Lijoi and Spano (2013) .

Example 6 *Suppose $\tilde{\mu}_1$ and $\tilde{\mu}_2$ are Gamma CRMs, i.e.*

$$\nu_i(dy) = e^{-y}y^{-1}dy, \quad i=1,2.$$

The bivariate Lévy intensity of a vector of Dirichlet process

$$\rho_2(y_1,y_2) = \left[\frac{1}{(y_1+y_2)^2}e^{-y_1-y_2} + \frac{1}{(y_1+y_2)}e^{-y_1-y_2}\right]\mathbb{1}_{(0,+\infty)^2}(y_1,y_2)$$

can be recovered by applying (1.11) *under the choice of the Lévy Copula*

$$C(y_1,y_2) = \Gamma(0,\Gamma^{-1}(0,y_1)+\Gamma^{-1}(0,y_2))$$

where $\Gamma(a,x) = \int_x^\infty s^{a-1}e^{-s}ds$ *is the incomplete gamma function and* $\Gamma^{-1}(a,x)$ *is its inverse function of* x*.*

1.2.2 Dependence through shared components

There are also other ways to introduce dependence without resorting to Lévy Copula. For example, a construction that does not rely on Lévy copula can be found in Lijoi, Nipoti and Prünster (2014a), where the dependence arises by virtue of a suitable construction of the Poisson random measures some of which are shared. Specifically, consider two identically distributed CRMs $\tilde{\mu}_1$ and $\tilde{\mu}_2$ with the same Lévy intensity ν. Lijoi, Nipoti and Prünster (2014a) show that $\tilde{\mu}_1$ and $\tilde{\mu}_2$ can be constructed by suitable mixtures of three independent CRMs μ_1, μ_2 and μ_0, whose Lévy intensities are respectively ν_1, ν_2 and ν_0.

$$\nu_0 = (1-z)\nu, \quad \nu_1 = \nu_2 = z\nu$$

for some random variable z taking values in $[0,1]$ and independent of μ_i, for $i = 0,1,2$. More precisely,

$$\tilde{\mu}_1 = \mu_1 + \mu_0, \quad \tilde{\mu}_2 = \mu_2 + \mu_0$$

Thus, the shared component μ_0 gives rise to the dependence between $\tilde{\mu}_1$ and $\tilde{\mu}_2$. The corresponding vector of random probability measures $(\tilde{p}_1, \tilde{p}_2)$ is constructed by

$$(\tilde{p}_1, \tilde{p}_2) = \left(\frac{\tilde{\mu}_1}{\tilde{\mu}_1(\mathbb{X})}, \frac{\tilde{\mu}_2}{\tilde{\mu}_2(\mathbb{X})} \right)$$

This representation is appealing because it allows the joint Laplace functional transform of $(\tilde{\mu}_1, \tilde{\mu}_2)$ having a simple structure, which is the key to derive the theoretical properties of some quantities of interest.

Also by considering the superposition of some shared components, Griffin, Kolossiatis and Steel (2013) introduce dependence in a similar way. Let $\boldsymbol{\mu} = (\mu_1, ..., \mu_p)$ be the vector of random measures with respective Lévy intensity ν_i, $i = 1, ..., p$. The random measures in the vector $\tilde{\boldsymbol{\mu}} = (\tilde{\mu}_1, ..., \tilde{\mu}_q)$ will be formed as

$$\tilde{\boldsymbol{\mu}} = D\boldsymbol{\mu}$$

where D is a $q \times p$ dimensional selection matrix (i.e. a matrix with only 0s and 1s as elements). Then $\tilde{\mu}_j$ is a Lévy process with Lévy intensity $\bar{\nu}_j = D_{j\cdot}\boldsymbol{\nu}$, where $D_{j\cdot}$ is

the j-th row of D and $\boldsymbol{\nu} = (\nu_1, ..., \nu_p)$. The vector of random probability measures $\tilde{\boldsymbol{p}} = (\tilde{p}_1, ..., \tilde{p}_q)$ is obtained by normalization of $\tilde{\boldsymbol{\mu}}$, i.e.

$$\tilde{p}_j = \frac{\tilde{\mu}_j}{\tilde{\mu}_j(\mathbb{X})}$$

1.2.3 Dependence through stick-breaking representation

Although the stick-breaking representation is out the scope of this thesis, it is worth to mention another way to construct dependence through the stick-breaking representation because of its computational merits. Let's go back to the stick-breaking representation of $\tilde{p}_i$ which states that $\tilde{p}_i = \sum_{j=1}^{\infty} W_{i,j}\delta_{X_{i,j}},\ i = 1, ..., d$, with $W_{i,1} = V_{i,1},\ W_{i,j} = V_{i,j}\prod_{l=1}^{j-1} V_{i,l}\ j \geq 2$ and $V_{i,j} \sim \text{Beta}(1, \alpha_i)$ and $X_{i,j}$ draws from some distribution G_i. For any $i = 1, ..., d$, $\tilde{p}_i$ is a $\text{DP}(\alpha_i, G_i)$. Dependence between any two random measures $\tilde{p}_i$ and $\tilde{p}_{i'}$ could be induced by the possible dependence between $V_{i,j}$ and $V_{i',j}$ or between $X_{i,j}$ and $X_{i',j}$. For instance, De Iorio et al. (2004) propose an ANOVA-type dependence for the law of the atoms, and later, Griffin and Steel (2006), define a class of DP with both dependent atoms and weights. The practical use of these models has been popular by the developments of suitable MCMC sampling algorithms, that make use of the stick-breaking representation of $\tilde{p}_i$, see Bassetti, Casarin and Leisen (2014). However, with this representation, it is not clear how to derive the analytical expression for some quantities of interest, such as EPPF and posterior distribution.

1.3 Species sampling models

One of the reasons for the popularity of DP models is the computational simplicity of posterior predictive inference. This simplicity is in part due to the almost sure discrete nature of a random probability measure with DP prior. Recall from the posterior predictive distribution of DP prior (1.2) and Poisson Dirichlet prior (1.3). In fact, the posterior predictive distribution provides another direction to generalize the DP models and develop new ones. We start this section by introducing species sampling sequences (SS sequence).

Definition 7 *A sequence of random variables $X_1, X_2, ...$, defined on a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ taking values in a Polish space, is a species sampling sequence (SS sequence), if the conditional distribution of X_{n+1} given $X(n) := (X_1, ..., X_n)$ has the*

following form:

$$P(X_{n+1} \in \cdot | X_1, \cdots, X_n) = \sum_{j=1}^{k_n} p_{n,j} \delta_{X_j^*}(\cdot) + r_{n,k_n+1} G_0(\cdot) \tag{1.13}$$

where k_n is the distinct number of values $(X_1^*, ..., X_{k_n}^*)$ among $(X_1, ..., X_n)$ and the non-negative weights $p_{n,j}$ need to satisfy $\sum_{j=1}^{k_n} p_{n,j} + r_{n,k_n+1} = 1$.

A species sampling sequence $(X_n)_{n\geq 1}$ can be interpreted as the sequential random sampling of individuals' species from a possibly infinite population of individuals belonging to several species. More specifically, X_1 is assigned a random tag, distributed according to $G_0(\cdot)$, as the species of the first individual. Given the tags $X_1, ..., X_n$ of the first n individuals observed, the species of the $(n+1)$th individual is a new species with probability r_{n,k_n+1} and it is equal to the observed species X_k with probability $\sum_{j=1}^{k_n} p_{n,j} \mathbb{1}_{\{X_j^* = X_k\}}$

If the SS sequence is an exchangeable sequence, then the weights $p_{n,j}$'s depend only on the cluster sizes. For instance, the Dirichlet process and two-parameter Poisson Dirichlet process represent two remarkable examples. More specifically, for the DP(α, G_0), $p_{n,j} = n_j/(n+\alpha)$ for $j = 1, ..., k_n$, with n_j the cluster size of the j-th cluster, $p_{n,k_n+1} = \alpha/(n+\alpha)$, and for the PD$(\sigma, \theta)$, $p_{n,j} = (n_j - \sigma)/(n+\theta)$ for $j = 1, ..., k_n$, $p_{n,k_n+1} = (\theta + \sigma k_n)/(n+\theta)$.

Whenever $p_{n,j}$'s does not only depend on the cluster sizes, the sequence $X_1, ..., X_n$ is no longer exchangeable. Exchangeability is a reasonable assumption in some clustering applications, but in many it is not. Consider data ordered in time, such as a time-stamped collection of news articles. In this setting, each article should tend to cluster with other articles that are nearby in time. Or, consider spatial data, such as pixels in an image or measurements at geographic locations. Here again, each datum should tend to cluster with other data that are nearby in space. While the traditional CRP mixture provides a flexible prior over partitions of the data, it cannot accommodate such non-exchangeability.

There has been an increasing focus on models which accommodate non-exchangeability in recent years. For example, Bassetti, Crimaldi and Leisen (2010) introduce a class of random sequences, called *generalized species sampling sequences*, which provides a new class of priors to address non-exchangeability. It has the following definition:

Definition 8 *A sequence $(X_n)_{n\geq 1}$ of random variables is a generalized species sampling sequence if:*

- X_1 *has distribution* G_0*;*

- *there exists a sequence* $(Y_n)_{n\geq 1}$ *of random variables such that, for each* $n \geq 1$*, the conditional distribution of* X_{n+1} *given* $X(n) = (X_1, ..., X_n)$ *and* $Y(n) = (Y_1, ..., Y_n)$ *is*

$$P\left(X_{n+1} \in \cdot | X(n), Y(n)\right) = \sum_{j=1}^{n} p_{n,j}(\pi^{(n)}, Y(n))\delta_{X_j^*}(\cdot) + r_n(\pi^{(n)}, Y(n))G_0(\cdot) \tag{1.14}$$

with $\pi^{(n)}$ *the random partition induced by* $X(n)$*;*

- X_{n+1} *and* $(Y_{n+j})_{j\geq 1}$ *are conditionally independent given* $X(n)$ *and* $Y(n)$*.*

Note that instead of describing the law of $(X_n)_{n\geq 1}$ with the sequence of the conditional distributions of X_{n+1} given $X(n)$, the generalized species sampling sequence has a latent process $(Y_n)_{n\geq 1}$ and it characterizes $(X_n)_{n\geq 1}$ with the sequences of the conditional distributions of X_{n+1} given $(X(n), Y(n))$.
Moreover, Bassetti, Crimaldi and Leisen (2010) illustrate with details two types of generalized species sampling sequences. One is the generalized Poisson-Dirichlet process, which is the conditionally identically distributed version of the well-known Poisson-Dirichlet process. The other class of sequences is what the authors called generalized Ottawa sequences (GOS).

Example 7 *(Generalized Poisson-Dirichlet process) Let* $\alpha \geq 0$ *and* $\theta \geq -\alpha$*. Consider the following sequences of functions:*

$$p_{n,i}(\pi^{(n)}, y(n)) := \frac{y_i - \alpha/n^{(i)}}{\theta + \sum_{j=1}^{n} y_j}, \quad i = 1, ..., n$$
$$r_n(\pi^{(n)}, y(n)) := \frac{\theta + \alpha k_n}{\theta + \sum_{j=1}^{n} y_j},$$

where $y(n) = (y_1, ..., y_n) \in [\alpha, +\infty)^n$*,* $n^{(i)}$ *is the size of the cluster which contains* i*-th observation,* k_n *is the number of clusters formed by the previous* n *observations* $(x_1, ..., x_n)$*. There exists a generalized species sampling sequence* $(X_n)_{n\geq 1}$ *for which*

$$P(X_{n+1} \in \cdot | X(n), Y(n)) = \sum_{j=1}^{k_n} \frac{(\sum_{i\in\pi^{(i)}} Y_i) - \alpha}{\theta + \sum_{i=1}^{n} y_i}\delta_{X_j^*}(\cdot) + \frac{\theta + \alpha k_n}{\theta + \sum_{i=1}^{n} y_i}G_0(\cdot), \tag{1.15}$$

where $(Y_n)_{n\geq 1}$ is a sequence of independent random variables such that each Y_n has law ν_n.

It is worthwhile to note that if $Y_n = 1$ for every $n \geq 1$, $\alpha \in [0, 1)$ and $\theta > -\alpha$, then we recover the well-known two parameters Poisson-Dirichlet process.

Example 8 *(Generalized Ottawa sequences) We say that a generalized species sampling sequence $(X_n)_{n\geq 1}$ is a generalized Ottawa sequence (GOS), if the following conditions hold for every $n \geq 1$.*

- *The sequences of functions $p_{n,i}(y(n))$ and $r_n(y(n))$ (for $i = 1, ..., n$) do not depend on the partition $\pi^{(n)}$.*
- *The functions r_n are strictly positive and*

$$r_n(Y_1, ..., Y_n) \geq r_{n+1}(Y_1, ..., Y_n, Y_{n+1})$$

  almost surely.

- *The function $p_{n,i}$ satisfy for each $y(n) = (y_1, ..., y_n)$, the equalities*

$$p_{n,i}(y(n)) = \frac{r_n(y(n))}{r_{n-1}(y(n-1))} p_{n-1,i}(y(n-1)), \text{ for } i = 1, ..., n-1$$
$$p_{n,n}(y(n)) = 1 - \frac{r_n(y(n))}{r_{n-1}(y(n-1))}$$

  with $r_0 = 1$.

We will only focus on one of novel GOS process introduced by Airoldi, Costa, et al. (2014), namely Beta-GOS process, where the weights are a product of independent Beta random variables.
We describe the Beta-GOS process by its PPF. Let $X(n) = (X_1, ..., X_n)$, the PPF is given by

$$P(X_{n+1} \in \cdot | X(n), W(n)) = \sum_{j=1}^{n} p_{n,j} \delta_{X_j}(\cdot) + r_n G_0(\cdot) \tag{1.16}$$

where the weights are defined by

$$p_{n,j} = (1 - W_j) \prod_{i=j+1}^{n} W_i, \quad r_{n,j} = \prod_{i=1}^{n} W_i$$

with $W(n) = (W_1, ..., W_n)$ being a sequence of independent Beta(α_n, β_n) random variables. The choice of Beta latent variables allows for a flexible specification of the species sampling weights, while retaining a simple and interpretable model together with computational simplicity. In fact, Airoldi, Costa, et al. (2014) have shown that Beta-GOS is a robust alternative to DP when the assumption of exchangeablity can hardly be applied, and also an alternative to customary Hidden Markov Model (HMM), especially when the number of states is unknown.
In section 3, we will introduce a new model based on the Beta-GOS sequence to model the neuroimage data, with the goal to accommodate both temporal and spatial dependency which are normally present in this kind of data.

1.4 Approximate Bayesian Computation(ABC) and its variant $\mathbf{BC}_{el}$

Many Bayesian applications involve likelihoods which are analytically infeasible or computationally expensive to evaluate. Such likelihoods naturally arise in many research areas, for example in Population Genetics (Beaumont et al. (2002), Drovandi and Pettitt (2010)), Epidemics (McKinley, Cook and Deardon (2009)) and Hidden Markov Models (Dean et al. (2014)).
An illustrative example would be a time series that can be characterized by a HMM. Due to the conditional dependencies between states at different time points, calculation of the likelihood of time series data is somewhat tedious, which illustrates the motivation to use ABC. See the following example for details.

Example 9 *(HMM) Let the random variable $X_t \in \mathcal{X}$ be the hidden state at time t, Y_t be the observation at time t. The conditional distribution of the hidden state is*

$$p(X_{t+1} = j | X_t = i) = p_{ij}$$

with initial conditions $\pi_0 = p(X_0 = i)$. The observation Y_t dependent on the current state X_t only, i.e.

$$p\left[Y_t = y | X(t), Y(t-1)\right] = p(Y_t = y | X_t)$$

where $X(t) = (X_1, ..., X_t)$ and $Y(t-1) = (Y_1, ..., Y_{t-1})$. The joint probability can be

described as

$$p[Y(t), X(t)] = \prod_{k=0}^{t} p(X_k|X_{k-1})p(Y_k|X_k)$$

To compute the likelihood function, we can use

$$p[Y(t)] = \sum_{X(t)\in\mathcal{X}^t} p[Y(t), X(t)]$$

However, as the number of possible sequences $X(t)$ is exponential, this direct computation becomes unfeasible unless t is small.

Another example would be the widely used mixed-effect models.

Example 10 *Let y_{ij} be the jth observation within subject (or blocking factor) i. Consider the following model of the form*

$$y_{ij} = x_{ij}\beta + b_i + \epsilon_{ij}$$

where x_{ij} are fixed effect predictors, $b_i \sim F$ and $\epsilon_{ij} \sim G$ with b_i and ϵ_{ij} uncorrelated. The likelihood is then

$$L(\beta) = \int \prod_{i,j} G(y_{ij} - x_{ij}\beta - b_i) d \prod_i F(b_i).$$

If G and F are not standard distributions as normal ones, the likelihood is extremely difficult to evaluate.

Approximate Bayesian Computational (ABC) methods allow us to manage situations where the likelihood is intractable. As remarked by Marin et al (2012), the first genuine ABC algorithm was introduced by Pritchard et al. (1999) in a Population Genetics setting. Precisely, suppose that the data $\boldsymbol{y} \in \mathcal{D} \subset \mathbb{R}^n$ is observed. Let $\varepsilon > 0$ be a tolerance level, η a summary statistic on $\mathcal{D}$ (which often is not sufficient) and ρ a distance on $\eta(\mathcal{D})$. Then, the algorithm works as follows

for $i = 1$ **to** N **do**
–**Repeat**
—-Generate θ' from the prior distribution $\pi(\cdot)$
—-Generate $\boldsymbol{z}$ from the likelihood $f(\cdot|\theta')$
–**until** $\rho\{\eta(\boldsymbol{z}), \eta(\boldsymbol{y})\} \leq \varepsilon$

set $\theta_i = \theta'$
end for

The basic idea behind the ABC is that, for a small (enough) ε and a representative summary statistic, we can obtain a reasonable approximation of the posterior distribution. Therefore, the choice of a summary statistics, a distance and a tolerance level play a crucial role to implement an efficient ABC algorithm. In order to relax some of the tuning problems, several recent papers have focused on strategies for setting the parameters of ABC algorithms, see for instance Fearnhead and Prangle (2012), Del Moral, Doucet and Jasra (2012) and Sisson et al. (2007).

1.4.1 Bayesian Computation with Empirical Likelihood

Recently there was a growing interest in methods where approximated likelihoods are used to deal with intractability. For example, Cabras, Nueda and Ruli (2015) apply quasi-likelihood in an ABC with the Markov chain Monte Carlo algorithm. Mengersen, Pudlo and Robert (2013) proposed an alternative approach that uses the well-established empirical likelihood approximation (BC_{el} sampler). In the latter one, the authors apply the method in a Bayesian framework to avoid the choice of the ABC parameters. The basic algorithm works in the following way: Firstly, generate M parameters θ_i from the prior distribution. Then, set the weight $\omega_i = L_{el}(\theta_i|y)$, where $L_{el}(\theta_i|y)$ is the empirical likelihood of θ_i given the observed data y. The output of BC_{el} is a sample of size M of parameters with associated weights, which operates as an importance sampling output. Details of this algorithm will be provided later in this section. However, the validation of the empirical likelihood depends on the choice of a set of constraints that ensures convergence.

The empirical likelihood has been developed by Owen (2010) as a non-parametric version of classical likelihood techniques. The main ideas of empirical likelihood can be shortly summarized as follows.

Let $X = (X_1, \ldots, X_n)^t$ be a random sample from an unknown probability function $f(\cdot)$. In practice, we observe $X_i = x_i$ (for $i = 1, \ldots, n$), where $x_1, \ldots, x_n$ are n known numbers. We will assume that f is a discrete distribution on $\boldsymbol{x} = (x_1, \ldots, x_n)$ with

$$p_i = f(x_i), \qquad i = 1, \cdots, n$$

where

$$p_i \geq 0 \qquad \sum_{i=1}^{n} p_i = 1$$

Since

$$P\left\{X_1 = x_1, \ldots, X_n = x_n\right\} = p_1 \cdot p_2 \cdots p_n$$

then the likelihood is

$$L\left(p_1, \ldots, p_n\right) \equiv L\left(p_1, \ldots, p_n; \boldsymbol{X}\right) = \prod_{i=1}^{n} p_i$$

$L\left(p_1, \ldots, p_n\right)$ is usually called the *empirical likelihood* function. The empirical likelihood approach defines the parameters as functionals of the distribution f, for instance as moments of f. Secondly, the empirical likelihood function is maximized subject to some constraints. Formally, the procedure can be described as follows.

$$L_{el}(\boldsymbol{\theta}|\boldsymbol{x}) = \max_{p_i} \prod_{i=1}^{n} p_i$$

subject to the constraints

$$\textstyle\sum_{i=1}^{n} p_i = 1; \quad \sum_{i=1}^{n} p_i h(x_i; \boldsymbol{\theta}) = 0; \quad p_i \geq 0$$

For example, in the one-dimensional case when $\theta = \mathbb{E}(X)$, the empirical likelihood in θ is the maximum of the product $p_1, ..., p_n$ under the constraint $p_1 x_1 + ... + p_n x_n = \theta$. In fact, this is an optimization problem with a set of constrains. In some cases, the choice of constrains can be challenging (see Owen (2010)).

Mengersen, Pudlo and Robert (2013) developed an alternative approach to standard ABC methods, based on the empirical likelihood approximation (BC_{el} sampler). The procedure can be summarized as follows

---

for $i = 1$ **to** M **do**

1. Generate θ_i from the prior distribution $\pi(\cdot)$
2. Set the weight $w_i = L_{el}(\theta_i|\boldsymbol{x})$

end for

---

where $L_{el}(\theta_i|\boldsymbol{x})$ is the empirical likelihood of θ_i given the observed data $\boldsymbol{x}$. The output of BC_{el} is a sample of size M of parameters with associated weights, which operate as an importance sampling output.

The main advantages of BC_{el}, when compared with standard ABC, are that they neither require simulations from the sampling model, nor any choice of parameters such as summary statistics, distance measure and tolerance. Bypassing model simulations sometimes leads to significant time savings in complex models, like those found in population genetics. However, BC_{el} still requires delicate calibrations in most cases. In this thesis we propose to replace the empirical likelihood in the BC_{el} sampler with a bootstrap likelihood approximation. The main motivation is that the choice of the constraints which ensures the convergence of the empirical likelihood is sometimes unclear.

1.5 Main results and contribution

Leisen and Lijoi (2011) introduced a bivariate vector of random probability measures with Poisson-Dirichlet marginals where the dependence is induced through a Lévy's Copula. In this thesis the same approach is used for generalizing such a vector to the multivariate setting. A first important contribution is the derivation of the Laplace functional transform which plays a crucial role in the derivation of the Exchangeable Partition Probability function (EPPF). Secondly, we provide an analytical expression of the EPPF for the multivariate setting. Finally, a novel Markov Chain Monte Carlo algorithm for evaluating the EPPF is introduced and tested. Besides, numerical illustrations of the clustering behaviour of the new prior are provided.

As stated in Section 1.3, species sampling models provide another perspective to Bayesian nonparametrics by describing the model with its posterior predictive distribution. The recent Beta-GOS model by Airoldi, Costa, et al. (2014) is a non-exchangeable species sampling sequences characterized by a tractable predictive probability function with weights driven by a sequence of independent Beta random variables. Our work aims to give a contribution by modifying the Beta-GOS process to accommodate both spatial and temporal correlations, which are present in neuroimage data like fMRI and EEG data. In these applications, the exchangeability is not an appropriate assumption. We believe that the new modified model is capable of detecting the activated voxels in the brain during a specific task,

that is, test whether a voxel exhibits neuronal activity in response to a stimulus or not at any time point. This is still an on-going work and the results we display in the thesis are partial results.

Our work is not limited to Bayesian nonparametrics and another part of the thesis is devoted to the development of approximate Bayesian computation (ABC) method, which is another popular area of nonparametric computational methods in Bayesian inference. As pointed out in the previous section, Mengersen, Pudlo and Robert (2013) proposed an alternative approach that uses the well-established empirical likelihood approximation (BC_{el} sampler), with the motivation to avoid the choice of the ABC parameters. In the same spirit, we propose an alternative algorithm BC_{bl}, which combines the original ABC idea with the use of bootstrap likelihood. The motivation is that the choice of the constraints which ensures the convergence of the empirical likelihood is sometimes unclear. The advantages of BC_{bl} are illustrated with several examples.

1.6 Outline

The structure of the thesis is as follows. In Chapter 2, we provide the multivariate extension of a vector of two-parameter Poisson-Dirichlet processes. In particular, in Section 2.1 some preliminaries definitions and results are presented. Moreover, the multivariate extension of the Leisen and Lijoi (2011) vector of Poisson-Dirichlet processes is introduced. Section 2.2 is devoted to the derivation of the multivariate Laplace exponent. Additionally, a result about the copula in (2.2) is given to explain the similarity with the Laplace exponent of the vector of Gamma processes introduced in Leisen, Lijoi and Spano (2013). In Section 2.3 an explicit expression of the EPPF is provided for the multivariate setting and, finally, in Section 2.4 an MCMC algorithm to evaluate the EPPF is introduced and used to give some information about the clustering behaviour of the new prior. In Chapter 3, we propose the Zero-inflated Beta-GOS process and its application to neuroimage data analysis. We describe the model in details (in section 3.1) and provide a MCMC posterior sampler (in section 3.3). We complete the chapter by presenting the results on simulated data. Chapter 4 focuses on the development of the Bayesian computation with bootstrap likelihood (BC_{bl}). More specifically, section 4.1 is devoted to the description of our methodology and in Section 4.2 the methodology is tested on several examples such as time series, population genetics, stochastic differential equations

and random fields.

Chapter 2

A multivariate extension of a vector of two-parameter Poisson-Dirichlet processes

The use of Bayesian non-parametric (BNP) priors in applied statistical modeling has become increasingly popular during the last few years. Since the paper of Ferguson (1973), the Dirichlet Process and their extensions have been used to address inferential problems in many fields. The increased interest in non-parametric Bayesian approaches to data analysis is motivated by a number of attractive inferential properties. For example, BNP priors are often used as flexible models to describe the heterogeneity of the population of interest, as they implicitly induce a clustering of the observations into homogeneous groups.

A very interesting property of the Dirichlet process is the discreteness of the distributions sampled from it, even when the base measure G_0 is continuous. This property is also quite obvious through the Polya urn representation. The *mixture of Dirichlet process* (MDP) model is based on the idea of constructing absolutely continuous random distribution functions and was first considered in Lo (1984). It models the distribution from which the x_i are drawn as a mixture of parametric distributions of the form $F(\cdot|\boldsymbol{\theta})$, with the mixing distribution over $\boldsymbol{\theta}$ being P. Let the prior for this mixing distribution be a Dirichlet process with scale parameter α

and base measure G_0, it yields the following model:

$$\begin{aligned} X_i \mid \boldsymbol{\theta}_i &\sim F(\boldsymbol{\theta}_i) \\ \boldsymbol{\theta}_i \mid P &\sim P \\ P &\sim \mathrm{DP}(\alpha, G_0). \end{aligned} \tag{2.1}$$

The most well known and widely applied MDP is the MDP with Gaussian kernel introduced in Lo (1984). This model is given by:

$$f_P(x) = \int \mathcal{N}(x|\mu, \sigma^2) dP(\boldsymbol{\theta}) \tag{3.1}$$

where $\mathcal{N}(x|\mu, \sigma^2)$ represents a normal density function, with $P \sim \mathrm{DP}(\alpha, G_0)$, a DP with parameters $\alpha > 0$, the scale parameter, and G_0, a distribution on $\mathbb{R} \times \mathbb{R}_+$ where $\boldsymbol{\theta} = (\mu, \sigma^2)$ with μ to represent the mean and σ^2 the variance of the normal component.

Despite its popularity, a simple MDP also has limitations. Recently, a growing literature in Bayesian non-parametrics proposed new priors for modelling situations where data may be divided into different groups. In this case, one would like to consider different densities for different groups instead of a single common density for all the data. For instance, Bassetti, Casarin and Leisen (2014) analyse the Business Cycle of the United States and the European Union with a Panel Var model where they assume a dependent BNP prior. As an illustration, consider this following special bivariate case of their model

$$\begin{pmatrix} Y_{1t} \\ Y_{2t} \end{pmatrix} = \begin{pmatrix} \mu_{1t} \\ \mu_{2t} \end{pmatrix} + \begin{pmatrix} Z_t' & O_{2p}' \\ O_{2p}' & Z_t' \end{pmatrix} \begin{pmatrix} \Upsilon_1 \\ \Upsilon_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{pmatrix}$$

for $t = 1, \ldots, T$, where $O_{2p} = (0, \ldots, 0)' \in \mathbb{R}^{2p}$, $\Upsilon_i = (\upsilon_{1,1,i}, \ldots, \upsilon_{1,2p,i})'$ and $Z_t = (Y_{1t-1}, \ldots, Y_{1t-p}, Y_{2t-1}, \ldots, Y_{2t-p})'$ and $\varepsilon_{it} \sim \mathcal{N}(0, \sigma_{it}^2)$ with ε_{1t} and ε_{2s} independent $\forall s, t$. Instead of considering Normal errors, they assume a dependent non-parametric prior on the parameters $(\mu_{it}, \sigma_{it}^2)$, $i = 1, 2$, with Poisson-Dirichlet process marginals. Roughly speaking, the errors are modeled with dependent infinite mixtures of Normal Distributions instead of single Normal Distributions. The main motivation of this approach is that it is able to capture many specific properties of time series, such as multimodality, skewness, excess of kurtosis and presence of outliers (e.g., see Griffin (2011) and Griffin and Steel (2011)) allowing information

pooling across series. The use of dependent Bayesian non-parametric priors is not confined only to time series and, after the seminal paper of MacEachern (1999), the problem of modelling a finite number of dependent densities has become an active area of research in Bayesian non-parametrics.
In this chapter, we extend some results presented in Leisen and Lijoi (2011) to the multidimensional case, precisely, the Laplace transform and the exchangeable partition probability function. We want to stress that the Laplace transform is the basis to prove theoretical results of the prior of interest. When the dimension is greater than two, the derivation of the Laplace transform is non-trivial and the result that we prove, is interesting compared with the Laplace transform of the n-dimensional vector of Dirichlet Processes studied in Leisen, Lijoi and Spano (2013). Indeed, both dependent priors have Laplace exponents that are the same function of the marginal Lévy-intensities. The reason is that the Lévy Copula behind both priors has the form

$$C(s_1, \ldots, s_d) = U(U^{-1}(s_1) + \cdots + U^{-1}(s_d)) \tag{2.2}$$

where the function U is the tail integral of the marginal Lévy intensity.
Finally, an MCMC algorithm for evaluating the EPPF is proposed and tested in some scenarios. Since the EPPF has not a closed form, this algorithm is an useful tool to compute the EPPF and the predictive distributions. The application of the algorithm to specific configurations allows some further considerations about the clustering behaviour of the new prior.
The results in this chapter have recently been published, see Zhu and Leisen (2014)

2.1 Preliminaries

Suppose $\tilde{\mu}_1, \ldots, \tilde{\mu}_d$ are completely random measures on $(\mathbb{X}, \mathcal{X})$ with respective marginal Lévy measures,

$$\bar{\nu}_i(\mathrm{d}x, \mathrm{d}y) = \alpha(\mathrm{d}x)\, \nu_i(\mathrm{d}y) \qquad i = 1, \ldots, d \tag{2.3}$$

The probability measure α on $\mathbb{X}$ is non–atomic (i.e. $\alpha(\{x\}) = 0$ for all $x \in \mathbb{X}$) and ν_i is a measure on $\mathbb{R}^+$ such that $\int_{\mathbb{R}^+} \min(y, 1)\, \nu_i(dy) < \infty$. We further suppose that $\tilde{\mu}_1, \ldots, \tilde{\mu}_d$ are *stable* CRMs, *i.e.*

$$\nu_i(dy) = \frac{\sigma}{\Gamma(1-\sigma)} y^{-1-\sigma}\, \mathrm{d}y \qquad 0 < \sigma < 1 \text{ and } i = 1, \ldots, d. \tag{2.4}$$

Moreover, $\tilde{\mu}_1, \ldots, \tilde{\mu}_d$ are dependent and the random vector $(\tilde{\mu}_1, \ldots, \tilde{\mu}_d)$ has independent increments, in the sense that for any A_1,...,A_n in $\mathcal{X}$, with $A_i \cap A_j = \emptyset$ for any $i \neq j$, the random vectors $(\tilde{\mu}_1(A_i), ..., \tilde{\mu}_d(A_i))$ and $(\tilde{\mu}_1(A_j), ..., \tilde{\mu}_d(A_j))$ are independent. This implies that for any set of measurable functions $\boldsymbol{f} = (f_1, \ldots, f_d)$ such that $f_i : \mathbb{X} \to \mathbb{R}^+$, $i = 1, \ldots, d$ are non-negative and $\int |f_i| \, d\tilde{\mu}_j < \infty$, one has a multivariate analogous of the Lévy-Khintchine representation of $(\tilde{\mu}_1, \ldots, \tilde{\mu}_d)$

$$\mathbb{E}\left[e^{-\tilde{\mu}_1(f_1)-\cdots-\tilde{\mu}_d(f_d)}\right] = \exp\left\{-\psi^*_{\rho,d}(\boldsymbol{f})\right\}. \tag{2.5}$$

where $\tilde{\mu}_i(f_i) = \int f_i d\tilde{\mu}_i$ and

$$\psi^*_{\rho,d}(\boldsymbol{f}) = \int_{\mathbb{X}} \int_{(0,\infty)^d} \left[1 - \mathrm{e}^{-y_1 f_1(x)-\cdots-y_d f_d(x)}\right] \, \rho_d(\mathrm{d}y_1, \ldots, \mathrm{d}y_d) \, \alpha(\mathrm{d}x). \tag{2.6}$$

An important issue is the definition of the measure ρ_d in (2.6): we will determine it in such a way that it satisfies the condition

$$\int_{(0,\infty)^{d-1}} \rho_d(\mathrm{d}x_1, \ldots, dx_{j-1}, A, dx_{j+1}, \ldots, dx_d) = \int_A \frac{\sigma}{\Gamma(1-\sigma)} y^{-1-\sigma} \, \mathrm{d}y$$

for any $j = 1, \ldots, d$ and $A \in \mathcal{B}(\mathbb{R}^+)$. In other words, the marginal Lévy intensities coincide with ν_i in (2.4). Recall from section 1.2.1, a Lévy Copula is a mathematical tool that allows to construct multivariate Lévy intensities with fixed marginals. Let $U_i(y) := \int_y^\infty \nu_i(s) \, \mathrm{d}s$ be the i–th marginal tail integral associated with ν_i. If both the copula C and the marginal tail integrals are sufficiently smooth, then

$$\rho_d(y_1, \ldots, y_d) = \frac{\partial^d C(s_1, \ldots, s_d)}{\partial s_1 \cdots \partial s_d} \bigg|_{s_1 = U_1(y_1), \ldots s_d = U_d(y_d)} \nu_1(y_1) \ldots \nu_d(y_d).$$

The Lévy-Clayton Copula is defined by

$$C_\gamma(s_1, \ldots, s_d) = (s_1^{-\gamma} + \cdots + s_d^{-\gamma})^{-\frac{1}{\gamma}} \qquad \gamma > 0.$$

where the parameter γ regulates the degree of dependence. Under the particular choice of stable marginals and $\gamma = \frac{1}{\sigma}$, then

$$\rho_d(y_1, \ldots, y_d) = \frac{(\sigma)_d}{\Gamma(1-\sigma)} |\boldsymbol{y}|^{-\sigma-d} \tag{2.7}$$

where $|\boldsymbol{y}| = y_1 + \cdots + y_d$ and $(\sigma)_d = \sigma(\sigma+1)\cdots(\sigma+d-1)$ is the ascending factorial.

If $d = 2$ then we recover the bivariate Lévy intensity used in Leisen and Lijoi (2011).

Let $(\tilde{\mu}_1, \ldots, \tilde{\mu}_d)$ be the vector of random probability measures defined in (2.5) with ρ_d as in (2.7). Suppose $\mathbb{P}_{i,\sigma}$ is the probability distribution of $\tilde{\mu}_i$, for $i = 1, \ldots, d$. Hence $\mathbb{P}_{i,\sigma}$ is supported by the space of all boundedly finite measures $\boldsymbol{M}_{\mathbb{X}}$ on $\mathbb{X}$ endowed with the Borel σ–algebra $\mathcal{M}_{\mathbb{X}}$ with respect to the $w^\sharp$–topology ("weak-hash" topology[1]). Introduce, now, another probability distribution $\mathbb{P}_{i,\sigma,\theta}$ on $(\boldsymbol{M}_{\mathbb{X}}, \mathcal{M}_{\mathbb{X}})$ such that $\mathbb{P}_{i,\sigma,\theta} \ll \mathbb{P}_{i,\sigma}$ and

$$\frac{\mathrm{d}\mathbb{P}_{i,\sigma,\theta}}{\mathrm{d}\mathbb{P}_{i,\sigma}}(\mu) = \frac{\Gamma(\theta)}{(K)^{\frac{1}{d}}}[\mu(\mathbb{X})]^{-\theta} \tag{2.8}$$

where

$$K = \int_{(0,\infty)^d} \left(\prod_{i=1}^{d} \lambda_i\right)^{\theta-1} e^{-\psi_{\rho,d}(\boldsymbol{\lambda})} d\boldsymbol{\lambda}$$

and $\psi_{\rho,d}(\boldsymbol{\lambda})$ is the Laplace exponent defined in formula (2.10). We denote with $\tilde{\mu}_{i,\sigma,\theta}$ a random element defined on $(\Omega, \mathcal{F}, \mathbb{P})$ and taking values in $(\boldsymbol{M}_{\mathbb{X}}, \mathcal{M}_{\mathbb{X}})$ whose probability distribution coincides with $\mathbb{P}_{i,\sigma,\theta}$. The random probability measure

$$\tilde{p}_i = \tilde{\mu}_{i,\sigma,\theta}/\tilde{\mu}_{i,\sigma,\theta}(\mathbb{X})$$

is a Poisson-Dirichlet process with parameter (σ, θ), see, e.g. Pitman and Yor (1997) and Pitman (2006), and the vector

$$(\tilde{p}_1, \ldots, \tilde{p}_d) \tag{2.9}$$

is a dependent vector of Poisson–Dirichlet random probability measures on $(\mathbb{X}, \mathcal{X})$. Note that, when $d = 2$, (2.9) coincides with the vector introduced in Leisen and Lijoi (2011).

Remark. Note that the change of measure in (2.8) differs from the one given in Leisen and Lijoi (2011). The latter contains a typo that slightly affects only the normalizing constant of the EPPF. Indeed, such a constant must coincide with $\frac{\Gamma(\theta)}{(K)^{\frac{1}{d}}}$.

[1]Recall that a sequence of measures $(m_i)_{i\geq 1}$ in $\boldsymbol{M}_{\mathbb{X}}$ converges, in the $w^\sharp$–topology, to a measure m in $\boldsymbol{M}_{\mathbb{X}}$ if and only if $m_i(A) \to m(A)$ for any bounded set $A \in \mathcal{X}$ such that $m(\partial A) = 0$. See Daley and Vere-Jones (2003) for further details.

2.2 The Laplace Exponent

In this section, the Laplace exponent of the vector $(\tilde{\mu}_1, \ldots, \tilde{\mu}_d)$ defined through the Lévy intensity in (2.7) is provided. This is an important tool for determining the Exchangeable Partition Probability Function in the next section.
Before getting started, it is worth noting that, since $(\tilde{\mu}_1, \ldots, \tilde{\mu}_d)$ has independent increments, its distribution is characterized by a choice of $f_1, \ldots, f_d$ in (2.5) such that $f_i = \lambda_i 1_A$ for any set A in $\mathcal{X}$, $\lambda_i \in \mathbb{R}^+$ and $i = 1, \ldots, d$. In this case

$$\psi^*_{\rho,d}(\boldsymbol{f}) = \alpha(A)\, \psi_{\rho,d}(\boldsymbol{\lambda})$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_d)$ and

$$\psi_{\rho,d}(\boldsymbol{\lambda}) = \int_{(\mathbb{R}^+)^d} \left[1 - \mathrm{e}^{-\langle \boldsymbol{\lambda}, \boldsymbol{y} \rangle}\right] \frac{(\sigma)_d}{\Gamma(1-\sigma)} |\boldsymbol{y}|^{-\sigma-d} \mathrm{d}\boldsymbol{y} \tag{2.10}$$

where $\boldsymbol{y} = (y_1, \ldots, y_d)$ and $\langle \boldsymbol{\lambda}, \boldsymbol{y} \rangle = \sum_{i=1}^d \lambda_i y_i$.
In Leisen and Lijoi (2011), the authors provide the expression of $\psi_{\rho,d}$ in the bidimensional case, i.e.

$$\psi_{\rho,2}(\lambda_1, \lambda_2) = \begin{cases} [\lambda_1^{\sigma+1} - \lambda_2^{\sigma+1}]/(\lambda_1 - \lambda_2) & \lambda_1 \neq \lambda_2 \\ (\sigma+1)\lambda_1^{\sigma} & \lambda_1 = \lambda_2 \end{cases} \tag{2.11}$$

In that scenario, the computation of $\psi_{\rho,2}$ is quite straightforward but it's not trivial in the multidimensional setting.

Proposition 1 *Let $\boldsymbol{\lambda} \in (\mathbb{R}^+)^d$ be a vector such that it consists of $l \leq d$ distinct values denoted as $\tilde{\boldsymbol{\lambda}} = (\tilde{\lambda}_1, \ldots, \tilde{\lambda}_l)$ with respective multiplicities $\boldsymbol{n} = (n_1, \ldots, n_l)$. Then*

$$\psi_{\rho,d}(\boldsymbol{\lambda}) = \psi_{\rho,d}(\tilde{\boldsymbol{\lambda}}, \boldsymbol{n}) = \left(\prod_{i=1}^{l} \frac{1}{\Gamma(n_i)} \frac{\partial^{n_i - 1}}{\partial^{n_i - 1} \tilde{\lambda}_i} \right) \left(\phi_l^{\sigma}(\tilde{\boldsymbol{\lambda}}) \prod_{i=1}^{l} \tilde{\lambda}_i^{n_i - 1} \right),$$

where

$$\phi_l^{\sigma}(\boldsymbol{x}) = \begin{cases} \sum_{i=1}^{l} \frac{x_i^{\sigma+l-1}}{\prod_{j=1, j\neq i}^{l} (x_i - x_j)} 1_{(x_1 \neq \ldots \neq x_l)} & \text{if } l > 1 \\ x_1^{\sigma} & \text{if } l = 1 \end{cases}$$

The proof of the result above can be found in the appendix. This result is interesting when compared with the Laplace exponent of the vector of Gamma Processes introduced in Leisen, Lijoi and Spano (2013). Indeed, both dependent priors have

Laplace exponents that are the same function of the marginal Lévy-intensities. The explanation is that the Lévy Copula behind the two processes has the same structure as showed in the following theorem.

Theorem 2 *Let ν be an univariate Lévy intensity such that*

$$\lim_{x\to\infty} x^i \nu^{(i-1)}(x) = 0 \tag{2.12}$$

where $v^{(i)}(x) = \frac{d^i}{dx^i}(\nu(x))$ and $\nu^{(0)}(x) = \nu(x)$. Let

$$\rho_d(y_1, ..., y_d) = (-1)^{d-1}\nu^{(d-1)}(y_1 + ... + y_d). \tag{2.13}$$

Hence, the Lévy intensity ρ_d can be recovered through the Lévy copula:

$$C(y_1, ..., y_d) = U(U^{-1}(y_1) + \cdots + U^{-1}(y_d))$$

where U is the tail integral of ν, i.e. $U(z) = \int_z^{+\infty} \nu(t)dt$, and U^{-1} its inverse.

Note that, the σ-stable and Gamma Lévy intensities satisfy condition (2.12). Moreover, the Lévy intensity introduced in (2.7) and the one introduced in Leisen, Lijoi and Spano (2013) have the form displayed in (2.13). Let $\Gamma(a, x) = \int_z^{+\infty} x^{a-1}e^x dx$ be the incomplete Gamma function, since

$$U(z) = \Gamma(0, z) \qquad \text{Gamma Process}$$

$$U(z) = \frac{1}{\Gamma(1-\sigma)} z^{-\sigma} \qquad \text{Stable Process}$$

it's straightforward to recover the copula introduced in Leisen, Lijoi and Spano (2013) and the Lévy-Clayton copula defined in (1.12) with $\gamma = \frac{1}{\sigma}$.

2.3 The Exchangeable Partition Probability Function

In this Section, the exchangeable partition probability function (EPPF) is computed for the vector defined in (2.9). As we will see, the expression of the EPPF depends

by a function g_ρ that is defined as

$$g_\rho(q_1,\ldots,q_d;\boldsymbol{\lambda}) = \int_{(0,\infty)^d} y_1^{q_1}\cdots y_d^{q_d} e^{-\psi_{\rho,d}(\boldsymbol{\lambda})}\rho_d(y_1,\ldots,y_d)d\boldsymbol{y} \tag{2.14}$$

An explicit expression of g_ρ can be found in the appendix. As in Leisen and Lijoi (2011), we are considering d groups of data with sample sizes $n_1,\ldots,n_d$ that are partial exchangeable, i.e.

$$\mathbb{P}\left[\boldsymbol{X}_{n_1}^{(1)} \in \times_{i_1=1}^{n_1} A_{i_1}^{(1)};\ldots;\boldsymbol{X}_{n_d}^{(d)} \in \times_{i_d=1}^{n_d} A_{i_d}^{(d)} \mid (\tilde{p}_1,\ldots,\tilde{p}_d)\right] = \prod_{i_1=1}^{n_1}\tilde{p}_1(A_{i_1}^{(1)})\times\cdots \\ \cdots\times\prod_{i_d=1}^{n_d}\tilde{p}_d(A_{i_d}^{(d)}).$$

with $\boldsymbol{X}_{n_i}^{(i)} = (X_1^{(i)},\ldots,X_{n_i}^{(i)})$, $i=1,\ldots,d$. This description of the model implies that the d samples $(X_1^{(1)},\ldots,X_{n_1}^{(1)}),\ldots,(X_1^{(d)},\ldots,X_{n_d}^{(d)})$ are independent conditional on $(\tilde{p}_1,\ldots,\tilde{p}_d)$. Given the discrete nature of the random probabilities in (2.9), there might be ties, i.e. common values among the samples $\boldsymbol{X}_{n_i}^{(i)}$, $i=1,\ldots,d$. Precisely, let $Z_1^*,\ldots,Z_K^*$ be the distinct values among the $(X_1^{(1)},\ldots,X_{n_1}^{(1)}),\ldots,(X_1^{(d)},\ldots,X_{n_d}^{(d)})$. Clearly, $1\leq K\leq n_1+\cdots+n_d$. Let $N_{i,j}$ be the number of $X^{(j)}$'s that are equal to Z_i^*, i.e.

$$N_{i,j} = \sum_{h=0}^{n_j} 1_{\{X_h^{(j)}=Z_i^*\}} \tag{2.15}$$

This means that the data can be described by the set

$$\{K, Z_1^*,\ldots,Z_K^*,(N_{1,1},\ldots,N_{K,1}),\ldots,(N_{1,d},\ldots,N_{K,d})\}$$

The *Exchangeable Partition Probability Function* (EPPF) is defined as

$$\Pi_k^{n_1,\ldots,n_d}(\boldsymbol{n_1},\ldots,\boldsymbol{n_d}) = \int_{\mathbb{X}}\mathbb{E}\left[\prod_{j=1}^{k}[\tilde{p}_1(dz_j)]^{n_{j,1}}\cdot\cdots\cdot[\tilde{p}_d(dz_j)]^{n_{j,d}}\right]$$

with $1\leq k\leq n_1+\cdots+n_d$ and for vector of non-negative integers such that $\sum_{h=0}^{k} n_{i,h} = n_i$. In the following theorem, an expression of the EPPF is provided.

Theorem 3 *For any positive integers $n_1,\ldots,n_d$ and k and vectors $\boldsymbol{n_i} = (n_{1,i},\ldots,n_{k,i})$ for $i=1,\ldots,d$ such that $\sum_{j=1}^{k} n_{j,i} = n_i$ and $n_{j,1}+\cdots+n_{j,d}\geq 1$, for $i=1,\ldots,d$, one*

has

$$\Pi_k^{n_1,\ldots,n_d}(\boldsymbol{n_1,\ldots,n_d}) = \frac{\displaystyle\int_{(0,\infty)^d} \lambda_1^{\theta+n_1-1}\cdots\lambda_d^{\theta+n_d-1}e^{-\psi_{\rho,d}(\boldsymbol{\lambda})}\times\prod_{j=1}^{k} g_\rho(n_{j,1},\ldots,n_{j,d};\boldsymbol{\lambda})d\boldsymbol{\lambda}}{K\prod_{i=1}^{d}(\theta)_{n_i}}$$

We close this section with some considerations about the symmetry of the EPPF. It worth to recall what happens in the univariate case. In this setting, we consider a sample of n observations $X_1,\ldots,X_n$ with distinct values $X_1^*,\ldots,X_k^*$ and frequencies $n_1,\ldots,n_k$ (n_i is the frequency of X_i^* among $X_1,\ldots,X_n$). The univariate EPPF of the Poisson-Dirichlet process is

$$\Pi_k^{(n)}(n_1,\ldots,n_k) = \frac{\prod_{i=1}^{k-1}(\theta+i\sigma)}{(\theta+1)_{n-1}}\prod_{j=1}^{k}(1-\sigma)_{n_j-1} \qquad 0<\sigma<1 \quad \theta>-\sigma \tag{2.16}$$

A nice feature about $\Pi_k^{(n)}$ is its symmetry: for any permutation τ of the integers $(1,\ldots,k)$ one has

$$\Pi_k^{(n)}(n_1,\ldots,n_k) = \Pi_k^{(n)}(n_{\tau(1)},\ldots,n_{\tau(k)}).$$

The invariance of the univariate EPPF with respect to permutations can be extended to the multivariate case. Consider the frequencies defined in equation (2.15) and let $\boldsymbol{r}_j = (n_{j,1},\ldots,n_{j,d})$. The partition probability function, seen as a function of $(\boldsymbol{r}_1,\ldots,\boldsymbol{r}_k)$, is symmetric in the sense that for any permutation τ of $(1,\ldots,k)$ one has

$$\Pi_k^{(n_1,\ldots,n_d)}(\boldsymbol{r}_1,\ldots,\boldsymbol{r}_k) = \Pi_k^{(n_1,\ldots,n_d)}(\boldsymbol{r}_{\tau(1)},\ldots,\boldsymbol{r}_{\tau(k)})$$

2.4 Numerical illustrations

The EPPF gives information about the clustering behaviour of the prior process. Moreover, the identification of the EPPF leads to the direct determination of the predictive distributions. For sake of illustration, consider the univariate two parameter Poisson-Dirichlet process discussed in the final part of the previous section and the EPPF displayed in (2.16). If $X_1,\ldots,X_n$ is a sample featuring $k\leq n$ dis-

tinct values $X_1^*, \ldots, X_k^*$ with respective frequencies $n_1, \ldots, n_k$ then the probability of sampling a new observation is

$$P(X_{n+1} = \text{new}|X_1, \ldots, X_n) = \frac{\Pi_{k+1}^{(n+1)}(n_1, \ldots, n_k, 1)}{\Pi_k^{(n)}(n_1, \ldots, n_k)} = \frac{\theta + \sigma k}{\theta + n}$$

and the probability of sampling an old observation X_j^*, $j = 1, \ldots, k$, is

$$P(X_{n+1} = X_j^*|X_1, \ldots, X_n) = \frac{\Pi_k^{(n+1)}(n_1, \ldots, n_j + 1, \ldots, n_k)}{\Pi_k^{(n)}(n_1, \ldots, n_k)} = \frac{1 - \sigma}{\theta + n}$$

The knowledge of the predictive distributions allows to implement Gibbs Sampling algorithms to address the posterior inference, see Escobar (1994), Escobar and West (1995) and Neal (2000). The two parameter Poisson-Dirichlet process illustrated above is a particular case in which the EPPF is known explicitly and, consequently, the predictive distributions are straighforward. In our case, the EPPF is not explicit since it depends by a multidimensional integral and then, the predictive distributions are unknown. However, accurate approximations of the EPPF could help to understand the clustering behaviour of the prior process and allow to compute the predictive distributions. For this reason, in this section we provide an automatic Markov Chain Monte Carlo (MCMC) algorithm for evaluating the EPPF numerically.

Algorithm 1 *Metropolis-Hastings for the EPPF*

Suppose that $\boldsymbol{\lambda}^{(t)} = (\lambda_1^{(t)}, \ldots, \lambda_d^{(t)})$

1. *Draw* $\boldsymbol{\lambda}' = (\lambda_1', \ldots, \lambda_d')$ *from a proposal distribution* $Q(\cdot|\boldsymbol{\lambda}^{(t)})$.

2. *Set* $\boldsymbol{\lambda}^{(t+1)} = \boldsymbol{\lambda}'$ *with probability*

$$\alpha(\boldsymbol{\lambda}^{(t)}, \boldsymbol{\lambda}') = \min\left\{1, \frac{\pi(\boldsymbol{\lambda}')Q(\boldsymbol{\lambda}^{(t)}|\boldsymbol{\lambda}')}{\pi(\boldsymbol{\lambda}^{(t)})Q(\boldsymbol{\lambda}'|\boldsymbol{\lambda}^{(t)})}\right\}$$

and $\boldsymbol{\lambda}^{(t+1)} = \boldsymbol{\lambda}^{(t)}$ *with probability* $1 - \alpha(\boldsymbol{\lambda}^{(t)}, \boldsymbol{\lambda}')$

Note that, the EPPF computed in Theorem 3 has the form

$$\Pi_k^{n_1,\ldots,n_d}(\boldsymbol{n_1},\ldots,\boldsymbol{n_d}) = \frac{\int F(\boldsymbol{\lambda})\pi(\boldsymbol{\lambda})d\boldsymbol{\lambda}}{K} \tag{2.17}$$

where

$$\pi(\boldsymbol{\lambda}) = e^{-\psi_{\rho,d}(\boldsymbol{\lambda})}\prod_{i=1}^{d}\lambda_i^{\theta-1}$$

and

$$F(\boldsymbol{\lambda}) = \left(\prod_{i=1}^{d}\frac{\lambda_i^{n_i}}{(\theta)_{n_i}}\right)\left(\prod_{j=1}^{k} g_\rho(n_{j,1},\ldots,n_{j,d};\boldsymbol{\lambda})\right)$$

The constant K acts naturally as the normalizing constant of the function $\pi(\boldsymbol{\lambda})$ and then, $\tilde{\pi} = \pi/K$ is a probability distribution. The Metropolis-Hastings algorithm allows to construct a Markov Chain $(X_i)_{i\geq 0}$ that has $\tilde{\pi}$ as stationary distribution and, for N large enough, an estimator of the integral (2.17) is

$$\Pi_k^{n_1,\ldots,n_d}(\boldsymbol{n_1},\ldots,\boldsymbol{n_d}) \cong \frac{1}{N}\sum_{i=1}^{N} F(X_i)$$

For a non-expert reader, the general steps (in our case) of the Metropolis-Hastings, are displayed in Algorithm 2. We suggest to set the proposal distribution as

$$Q(\cdot|\boldsymbol{\lambda}^{(t)}) = q_1(\cdot|\lambda_1^{(t)})\cdots q_d(\cdot|\lambda_d^{(t)})$$

Since θ regulates the shape of the proposal distribution, we need to choose $q_i(\cdot|\lambda_i^{(t)})$ according to its value. If $\theta \leq 1$ then $q_i(\cdot|\lambda_i^{(t)})$, $i = 1,\ldots,d$ is a Weibull Distribution, otherwise if $\theta > 1$, then $q_i(\cdot|\lambda_i^{(t)})$, $i = 1,\ldots,d$, is the density of a truncated Normal distribution in the interval $[0,\infty)$ with mean $\lambda_i^{(t)}$ and standard deviation s_i. When $\theta > 1$, a guideline for setting the standard deviations $\boldsymbol{s} = (s_1,\ldots,s_d)$ could be

$$\boldsymbol{s} = \text{Argmax } \pi(\boldsymbol{\lambda})$$

On the other hand, when $\theta \leq 1$, a guideline could be to set the shape parameter of the Weibull distribution equal to σ. Since the heaviness of the tails increases as σ decreases, we suggest the following empirical rule for the scale parameter: if $\sigma \geq 0.5$ then set it equal to 1 otherwise, set it at least 10. In the set of experiments, we run 20 chains for 20000 iterations with a burn in period of 5000 iterations. First of all,

		Numerical Integrator		MCMC	
θ	σ	$k=1$	$k=2$	$k=1$	$k=2$
0.5	0.25	0.294	0.706	0.2985 ± 0.0033	0.7075 ± 0.2022
0.5	0.5	0.1894	0.8106	0.1897 ± 0.0013	0.7947 ± 0.0158
0.5	0.8	0.0729	0.927	0.0729 ± 0.0005	0.9150 ± 0.0142
1	0.25	0.1862	0.8138	0.1867 ± 0.0034	0.8096 ± 0.0036
1	0.5	0.1215	0.8785	0.1218 ± 0.0008	0.8818 ± 0.0091
1	0.8	0.0475	0.9525	0.0476 ± 0.0003	0.9521 ± 0.0130
3	0.25	0.0734	0.9053	0.0759 ± 0.0048	0.9154 ± 0.0882
3	0.5	0.0496	0.9504	0.0498 ± 0.0012	0.9397 ± 0.0486
3	0.8	0.0197	0.981	0.0198 ± 0.0003	0.9769 ± 0.050
5	0.25	**0.0423**	**0.8585**	0.0472 ± 0.0004	0.9593 ± 0.0327
5	0.5	0.0303	0.9439	0.0313 ± 0.0003	0.9577 ± 0.0252
5	0.8	**0.0126**	**1.0061**	0.0124 ± 0.0001	0.9748 ± 0.0201

Table 2.1: Matlab integrator vs MCMC assuming $n_1 = 1$ and $n_2 = 1$

we focus on the bidimensional case to test the performance of our MCMC algorithm. Note that the expression of the EPPF in Theorem 3 is done by 2 integrals, one at the numerator and one at the denominator (i.e. the constant K). When $d = 2$, in a similar fashion of Leisen and Lijoi (2011), both integrals can be reduced to one dimensional integrals in the interval $[0, 1]$ and evaluated accurately through the standard one-dimensional matlab integrator. This allows a comparison with our MCMC algorithm and the results are displayed in Table 1 for different values of θ and σ.

As we can see in Table 2.1, our algorithm has a good level of accuracy compared with the matlab integrator. In some cases, it performs better, i.e when $(\theta = 5, \sigma = 0.25)$ and $(\theta = 5, \sigma = 0.8)$. This inaccuracy of the matlab integrator is due to the explosion of the values of both integrals at the numerator and denominator. Our algorithm, is immune to this problem as well as dimensional problems. In a slightly more advanced case, the same comparison is done when $n_1 = 2$ and $n_2 = 1$, see Table 2.2. Although these examples are very simple, we can do some comments about the clustering behaviour of such a vector. Precisely, if θ or σ increases then the number of clusters increases. Moreover, in table 2 it can be observed that when $\theta = 0.5$, the probability of the configuration with k=1 cluster changes its trend as σ increases, compared with the first 2 configurations with k=2 clusters. Indeed, when $\sigma = 0.25$, the first one has higher probability than the second one and, the exact opposite is true when $\sigma = 0.8$. This suggests that an higher σ encourages an higher number of clusters while a lower sigma suppresses the creation of new clusters.

Anyway, for a better understanding of the qualitative behaviour, we need to test the prior on more sophisticated configurations.

		Numerical Integrator			MCMC		
θ	σ	$k=1$	$k=2$	$k=3$	$k=1$	$k=2$	$k=3$
0.5	0.25	0.1715	0.1225	0.353	0.1700 ± 0.0043	0.1227 ± 0.0069	0.3503 ± 0.0295
			0.1225			0.1227 ± 0.0069	
			0.2295			0.2285 ± 0.0124	
0.5	0.5	0.0947	0.0947	0.5404	0.095 ± 0.0016	0.0959 ± 0.004	0.5478 ± 0.0764
			0.0947			0.0959 ± 0.004	
			0.1755			0.1753 ± 0.0117	
0.5	0.8	0.0292	0.0438	0.8034	0.0289 ± 0.0004	0.0439 ± 0.0013	0.8037 ± 0.0341
			0.0438			0.0439 ± 0.0013	
			0.0789			0.0795 ± 0.0023	
3	0.25	0.0161	0.0574	0.7355	0.0162 ± 0.0004	0.0570 ± 0.0033	0.7331 ± 0.0798
			0.0574			0.0570 ± 0.0033	
			0.1124			0.1136 ± 0.0072	
3	0.5	0.0093	0.0403	0.8316	0.0092 ± 0.0001	0.0398 ± 0.0013	0.8260 ± 0.07
			0.0403			0.0398 ± 0.0013	
			0.0785			0.0769 ± 0.0025	
3	0.8	0.0015	0.0167	0.9319	0.0029 ± 0.00001	0.0164 ± 0.001	0.9154 ± 0.0787
			0.0167			0.0164 ± 0.001	
			0.0282			0.0314 ± 0.0007	

Table 2.2: Matlab integrator vs MCMC assuming $n_1 = 2$ and $n_2 = 1$

We consider the three dimensional case where $n_1 = 40, n_2 = 20$ and $n_3 = 30$. In Figure 2.1 it is displayed the log scale EPPF for two different clustering behaviours given different values of θ and σ. The left hand side is the plot of log EPPF of 1 cluster case and the right hand side is the one with a configuration of k=3 clusters and multiplicities $\boldsymbol{n_1} = (10, 0, 0), \boldsymbol{n_2} = (10, 10, 10)$ and $\boldsymbol{n_3} = (20, 10, 20)$.
The probability of k=1 cluster decreases as σ increases and also together with the increase of θ, which evidently suggests that lower σ and lower θ depresses the creation of new clusters. In addition, the comparison of the probabilities of the two cases implies that the first case, which has only 1 cluster, happens with a enormously bigger chance than the second one. Similar trends has been observed on higher dimensional cases.

2.5 Conclusions

In this chapter we extended the bivariate vector of Leisen and Lijoi (2011) to the multivariate setting. The specification of the vector through completely random measures allows to determine the Laplace functional transform which is the basis to derive some quantities of interest. Thus, a first contribution deals with the derivation of the Laplace transform which is non-trivial in the multivariate setting. In a similar fashion of Leisen and Lijoi (2011), the knowledge of the Laplace transform allows to provide, as a second contribution, an expression of the Exchangeable Partition

Log EPPF in three dimensions

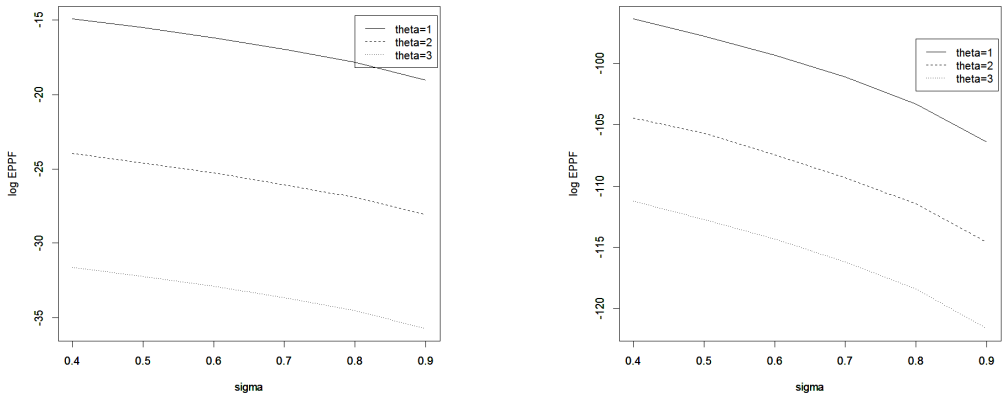

Figure 2.1: Left column: the log EPPF of k=1 cluster. Right column: the log EPPF of designated k=3 clusters

Probability Function (EPPF). Finally, a new MCMC algorithm has been introduced for evaluating the EPPF. The performance of the algorithm are tested as well as the clustering behavior of the new prior.

Chapter 3

Zero-inflated Beta-GOS process and its application to neuroimage data analysis

Statistical methods play a crucial role in the analysis of functional magnetic resonance imaging (fMRI) data, due to their complex spatial and temporal correlation structure. In fMRI experiments, neuronal activation in response to an input stimulus occurs in milliseconds and it is not observed directly. Instead, the blood oxygenation level dependent (BOLD) signal contrast is measured on the entire brain, since the metabolic process increases blood flow and volume in the activated areas following neuronal activation. As a brain area becomes active, e.g., in response to a task, there is an increase in local oxygen consumption and, consequently, more oxygen-rich blood flows to the active brain area.

Common approaches to the analysis of such data would calculate voxel-wise t-test or ANOVA statistics and/or fit a linear model at each voxel after a series of preprocessing steps, including scanner drift and motion corrections, adjustment for cardiac and respiratory-related noise, normalization and spatial smoothing, see Huettel, Song and McCarthy (2004). However, spatial correlation is expected in a voxel-level analysis of fMRI data because the response at a particular voxel is likely to be similar to the responses of neighboring voxels. Also, correlation among voxels does not necessarily decay with distance. These features of the data make single-voxel approaches not appropriate, as the test statistics across voxels are not independent. In addition, serious multiplicity issues arise, due to the large dimensionality of the data.

In this chapter, we propose a novel Bayesian nonparametric approach for modeling brain connectivity. More specifically, our goal is to provide a model that allows us to test whether a voxel exhibits neuronal activity in response to a stimulus or not at any time point t. Our proposed model is a modification of Beta-GOS process introduced by Airoldi, Costa, et al. (2014) (see (1.16)).
This is still an ongoing work with Michele Guindani (M.D. Anderson Cancer Center), Marina Vannucci (Rice University), Alberto Cassese (Rice University) and Fabrizio Leisen (University of Kent). Results displayed in this chapter are partial results.

3.1 Model

In an fMRI experiment, the whole brain is scanned at multiple time points and a time series of BOLD response is acquired for each voxel of the brain while the subject performs a set of tasks. Let $Y_{t,\nu}$ be the $T \times V$ matrix of response data, where $t = 1, ..., T$ indicates the time points and $\nu = 1, ..., V$ indicates voxels in the brain. We assume the data can be described as

$$Y_{t,\nu}|\mu_{t,\nu} \overset{ind}{\sim} p(y_{t,\nu}|\mu_{t,\nu}),$$

where $p(\cdot|\mu_{t,\nu})$ is some probability density.
We are firstly going to describe the model, which we call zero-inflated Beta-GOS process, and then visualize its application to neuroimaging data. We call a sequence of random variable $(X_n)_{n\geqslant 1}$ a zero-inflated Beta-GOS process if it can be characterized by the following predictive distribution

$$P\{X_{n+1} \in \cdot|\boldsymbol{X}(n), \boldsymbol{W}(n)\} = \omega_{n+1}\left\{\sum_{j=1}^{n} p_{n,j}\delta_{X_j}(\cdot) + r_n G_0(\cdot)\right\} + (1-\omega_{n+1})\delta_0(\cdot), \tag{3.1}$$

where the weights are defined by

$$p_{n,j} = (1-W_j)\prod_{i=j+1}^{n} W_i, \quad r_n = \prod_{i=1}^{n} W_i, \quad \omega_{n+1} \in (0,1)$$

with $W(n) = (W_1, ..., W_n)$ a vector of independent beta random variables W_k taking values from Beta(α_k,β_k), and δ_x is the Dirac function centered at x. The zero-inflated Beta-GOS process can be seen as a mixture of Beta-GOS process (1.16) and a Dirac

distribution at 0. In fact, X_{n+1} follows a Beta-GOS process with probability ω_{n+1}, or will be directly assigned the value 0 with probability $(1-\omega_{n+1})$. This setting ensures that the zero-inflated Beta-GOS process almost surely puts some positive mass on 0. More precisely, X_{n+1} will be directly assigned the value 0 with probability $1-\omega_{n+1}$; the probability of paring X_{n+1} to one of the previous $X_j,\ j=1,...,n$ will be $\omega_{n+1}\cdot p_{n,j}$; X_{n+1} will result in a new value from G_0 with probability $\omega_{n+1}\cdot r_n$. Notice that if X_{n+1} takes the value 0, other than firstly being assigned directly to value 0, it could also originate from one of the previous $X_k=0,\ k\in\{1,2,...,n\}$.

Going back to the fMRI data, for each voxel ν, we assume an underlying zero-inflated Beta-GOS process, i.e. we assume that $\boldsymbol{\mu}_\nu=(\mu_{1,\nu},...,\mu_{T,\nu})^T$ is a zero-inflated Beta-GOS process with parameters $\boldsymbol{\alpha}_\nu=(\alpha_{1,\nu},...,\alpha_{T,\nu})^T$, $\boldsymbol{\beta}_\nu=(\beta_{1,\nu},...,\beta_{T,\nu})^T$, $\boldsymbol{\omega}_\nu=(\omega_{1,\nu},...,\omega_{T,\nu})^T$ and base measure G_0. The predictive probability can thus be described as

$$P\{\mu_{t+1,\nu}\in\cdot|\boldsymbol{\mu}_\nu(t),\boldsymbol{W}_\nu(t)\}=\omega_{t+1,\nu}\left\{\sum_{j=1}^{t}p_{t,j,\nu}\delta_{\mu_{j,\nu}}(\cdot)+r_{t,\nu}G_0(\cdot)\right\}+(1-\omega_{t+1,\nu})\delta_0(\cdot) \tag{3.2}$$

Note that, conditional on $\mu_{t+1,\nu}$ not being equal to one of the previous $\mu_{i,\nu},\ i=1,...,t$, then $\mu_{t+1,\nu}$ will be set to zero with probability $1-\omega_{t+1,\nu}$ or will be sampled from G_0 with probability $\omega_{t+1,\nu}$.

The Beta-GOS process well captures the time dependency of neuroimaging data, but it does not take into account the spatial dependency. In the brain network, neighboring voxels often tend to be activated or inactivated together. In order to model this feature, it is helpful to bring in a binary random variable $\gamma_{t,\nu}$ which identifies whether $\mu_{t,\nu}$ is zero or not. In other words, we set $\gamma_{t,\nu}=0$ if $\mu_{t,\nu}=0$ and $\gamma_{t,\nu}=1$ if $\mu_{t,\nu}\neq 0$. A Markov random field (MRF) model is then placed on the parameter $\boldsymbol{\omega}_\nu$ to explicitly account for the spatial correlation. Specifically, $\omega_{t,\nu}$ can be expressed by

$$\omega_{t,\nu}=\frac{\exp(d+e\sum_{k\in N_{t,\nu}}\gamma_k)}{1+\exp(d+e\sum_{k\in N_{t,\nu}}\gamma_k)} \tag{3.3}$$

where $N_{t,\nu}$ is the set of neighboring voxels of voxel ν at time point t, the parameter $d\in(-\infty,+\infty)$ represents the expected prior number of active voxels, and controls the sparsity, while $e>0$ is a smoothing parameter. Accordingly, $\omega_{t,\nu}$ increases if

more of the neighboring voxels are activated, which also results in a greater probability of being activated for the voxel ν at time point t.

3.2 MCMC posterior sampling

Traditional Bayesian nonparametric algorithms use cluster labels as an indicator of the cluster membership for each observation. Airoldi, Costa, et al. (2014) use a different sequence of labels $(C_n)_{n\geq 1}$ which records the pairing of each observation. Precisely, $C_i = j$ means that among those observations with index $j < i$, the i-th observation has been matched to the j-th one. $C_i = i$ suggests that the i-th observation is the starter of a new cluster. We adopt the same pairing labels $(C_{t,\nu})_{t\geq 1}$ recording the paring of each observation according to 3.2 for the voxel ν. Clearly, for the first observation $C_{1,\nu} = 1$; $\gamma_{1,\nu} = 1$ with probability $\omega_{1,\nu}$ and $\gamma_{1,\nu} = 0$ otherwise. Note that if $C_{t,\nu} = j$, it means that the t-th observation is assigned to the same cluster as the j-th one, hence $\gamma_{t,\nu} = \gamma_{j,\nu}$. In particular, if the t-th observation is not paired to any of those preceding, $C_{t,\nu} = t$; in this case, the t-th point could possibly consist of a draw from the base distribution G_0 which result in $\gamma_{t,\nu} = 1$, or consist of value 0 from the zero point mass δ_0, in which case $\gamma_{t,\nu} = 0$. These pairing labels are useful to develop an MCMC sampling scheme for non-exchangeable processes.

For any $t \leq T$, let $\boldsymbol{C}_\nu(T) = (C_{1,\nu}, ..., C_{T,\nu})$ be the vector of pairing labels, and $\boldsymbol{W}_\nu(T) = (W_{1,\nu}, ..., W_{T,\nu})$ be the vectors of W elements. It is straightforward to see that the pairing sequence $(C_{t,\nu})_{t\geq 1}$ has the following distribution

$$\begin{aligned} &P\{C_{t,\nu} = j|\boldsymbol{C}_\nu(t-1), \boldsymbol{W}_\nu(t-1)\} = P\{C_{t,\nu} = j|W_{1,\nu}, ..., W_{t-1,\nu}\} \\ =&\omega_{t,\nu} \cdot p_{t-1,j,\nu}\mathbb{1}\{j \neq t\} + (\omega_{t,\nu} \cdot r_{t-1,j,\nu} + 1 - \omega_{t,\nu})\mathbb{1}\{j = t\} \end{aligned} \tag{3.4}$$

where $\mathbb{1}(\cdot)$ is the indicator function. Let $\boldsymbol{C}_{-t,\nu} = (C_{1,\nu}, ..., C_{t-1,\nu}, C_{t+1,\nu}, ..., C_{T,\nu})$ be the same vector as $\boldsymbol{C}_\nu(T)$ where the t-th element has been removed. Similarly, let $\boldsymbol{\gamma}_\nu(T) = (\gamma_{1,\nu}, ..., \gamma_{T,\nu})$, $\boldsymbol{\gamma}_{-t,\nu} = (\gamma_{1,\nu}, ..., \gamma_{t-1,\nu}, \gamma_{t+1,\nu}, ..., \gamma_{T,\nu})$ and $\boldsymbol{W}_{-t,\nu} = (W_{1,\nu}, ..., W_{t-1,\nu}, W_{t+1,\nu}, ..., W_{T,\nu})$ be the previous corresponding vectors where the t-th element has been removed. The full conditionals of the pairing label $C_{t,\nu}$ is

given by

$$P\{C_{t,\nu} = j|\boldsymbol{C}_{-t,\nu}, \boldsymbol{\gamma}_{-t,\nu}, \boldsymbol{Y}_\nu(T), \boldsymbol{W}_\nu(T)\}$$
$$\propto P\{\boldsymbol{Y}_\nu(T)|C_{t,\nu} = j, \boldsymbol{C}_{-t,\nu}, \boldsymbol{\gamma}_\nu(T), \boldsymbol{W}_\nu(T)\} P\{C_{t,\nu} = j|\boldsymbol{C}_{-t,\nu}, \boldsymbol{\gamma}_{-t,\nu}, \boldsymbol{W}_\nu(T)\} \tag{3.5}$$

The second term of (3.5) is given by (3.4). We shall see the first term in the following. Let $\Pi(\boldsymbol{C}_{-t,\nu}, j)$ denotes the partition generated by $(C_{1,\nu}, ..., C_{i-1,\nu}, j, C_{i+1,\nu}, ..., C_{T,\nu})$. We furthermore divide the partition into two groups based on whether the value of the cluster is zero or not, that is, $\Pi(\boldsymbol{C}_{-t,\nu}, j) = \{\Pi_0(\boldsymbol{C}_{-t,\nu}, j), \Pi_1(\boldsymbol{C}_{-t,\nu}, j)\}$, where $\Pi_0(\boldsymbol{C}_{-t,\nu}, j)$ represents the cluster of zero-valued $\mu_{t,\nu}$ and $\Pi_1(\boldsymbol{C}_{-t,\nu}, j)$ the remaining nonzero-valued cluster. So the first term of (3.5) is

$$P\{\boldsymbol{Y}_\nu(T)|C_{t,\nu} = j, \boldsymbol{C}_{-t,\nu}, \boldsymbol{\gamma}_\nu(T), \boldsymbol{W}_\nu(T)\} =$$
$$\Big\{ \prod_{l \in \prod_0(\boldsymbol{C}_{-t,\nu}, j)} p(Y_{l,\nu}|0) \Big\} \prod_{k=1}^{|\Pi_1(\boldsymbol{C}_{-t,\nu}, j)|} \int \prod_{l \in \prod_1(\boldsymbol{C}_{-t,\nu}, j)_k} p(Y_{l,\nu}|\mu^*_{j,\nu}) G_0(d\mu^*_{j,\nu}) \tag{3.6}$$

The further details related to the above expression can be found in the Appendix. With regard to the full conditional for the variables $\boldsymbol{W}_\nu(T)$, Airoldi, Costa, et al. (2014) has shown that

$$W_{t,\nu}|\boldsymbol{C}_\nu(T), \boldsymbol{W}_{-t,\nu}, \boldsymbol{Y}_\nu(T) \sim \text{Beta}(A_{t,\nu}, B_{t,\nu}), \tag{3.7}$$

where $A_{t,\nu} = \alpha_{t,\nu} + \sum_{j=t+1}^{T} \mathbb{1}\{C_j < t \text{ or } C_j = j\}$ and $B_{t,\nu} = \beta_{t,\nu} + \sum_{j=t+1}^{T} \mathbb{1}\{C_j = t\}$.

As for the cluster centroids $\mu^*_{i,\nu}$'s, the algorithm described above actually integrates over the distribution of the $\mu^*_{i,\nu}$ to achieve faster mixing of the chain. However, it is possible to sample the unique values at each iteration of the Gibbs sampler. More specifically, in the 0-valued cluster, $\mu^*_{j,\nu} = 0$; in the nonzero-valued clusters,

$$P\{\mu^*_{j,\nu}|\boldsymbol{C}_\nu(T), \boldsymbol{W}_\nu(T), \boldsymbol{Y}_\nu(T)\} = \prod_{i \in \Pi_{j,\nu}(T)} p(Y_{i,\nu}|\mu^*_{j,\nu}) G_0(d\mu^*_{j,\nu})$$

where $\Pi_{j,\nu}(T)$ denotes the non-zero partition set consisting of $\mu_{i,\nu} = \mu^*_{j,\nu},\ i = 1, ..., T$.

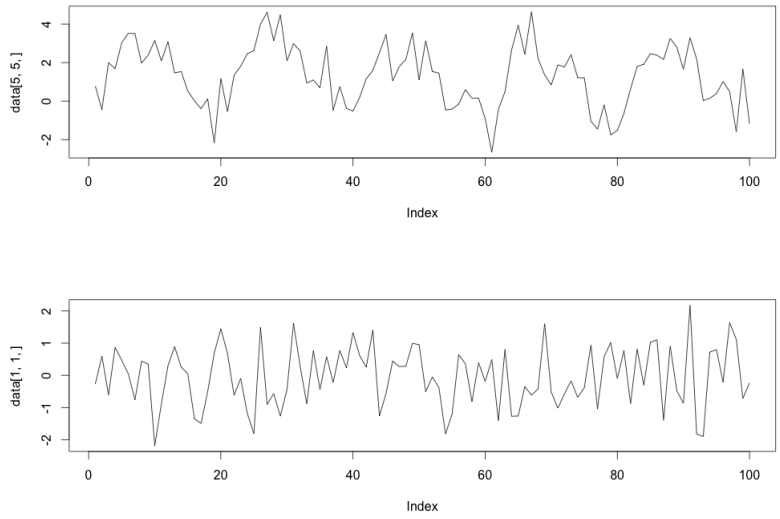

Figure 3.1: An example of signals of an activated voxel (above) and an inactivated one (below). Each activation block lasts 10 time points and the first starts from $t = 4$, followed by an non-activation block that also lasts 10 time points. Then both states appear alternately. The SNR is 1.5.

3.3 Simulation Study

In this section, we apply our model on simulated fMRI data to perform hypothesis test on whether the voxles are activated or not at each time point.

In the simulations presented below we consider $T = 100$ images of 10×10 voxels. The data is generated using R package **neuRosim**. The block design consists of two different conditions, activated and non-activated, with the active pattern displayed in Figure 3.2. We simulate the time series with signal-to-noise ratio (SNR) set as 1.5 and the type of noise as white. Figure 3.1 shows an example of signals of an activated voxel and an inactivated voxel.

3.3.1 Model specification

We specify the model (3.2) as follows. First, we assume a Gaussian distribution for the observables, $Y_{t,\nu} \sim \text{Normal}(\mu_{t,\nu}, \tau^2)$. Assume that the vector $(\mu_{1,\nu}, ..., \mu_{T,\nu})$ follows an zero-inflated Beta-GOS process. G_0 is assumed to be normal, $G_0 = \text{Normal}(\mu_0, \sigma_0^2)$, and $\tau^2 \sim \text{Inv-Gamma}(a_0, b_0)$. The parameters of the latent Beta

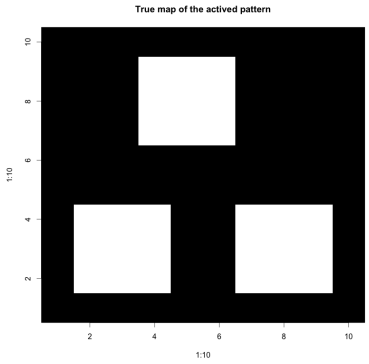

Figure 3.2: True activation map with three activated blocks

reinforcements, $W_{t,\nu} \sim \text{Beta}(\alpha_{t,\nu}, \beta_{t,\nu})$, are updated in each iteration of Gibbs sampler.

We consider two cases of Beta hyper parameters $\alpha_{t,\nu} = 1, \beta_{t,\nu} = 1$ and $\alpha_{t,\nu} = t, \beta_{t,\nu} = 1$. Furthermore we assume $G_0 = \text{Normal}(0, 10^2)$ to allow a relatively large range for the proposed value of μ_i, and $\tau^2 \sim$ Inv-Gamma(3165,1780) in the model fitting. This choice of the Inverse-Gamma hyper-parameters allows τ to have mean around 0.75 and reasonable variability. As for the parameters of Gaussian random field prior, we select $d = -3$, $e = 0.1$. We will discuss the sensitivity to these choices below.

At each MCMC iteration, we update the paring label $\boldsymbol{C}_\nu(T)$ according to 3.5 and $\boldsymbol{\gamma}_\nu(T)$ for all the voxels. We run the MCMC chains with 1000 iterations, discarding the first 300 ones as burn-in.The posterior activation probability maps are obtained by setting the posterior probability threshold at 0.8, that is, an individual voxel ν at a time point t is categorized as active if the posterior probability $p(\gamma_{t,\nu}|\boldsymbol{Y}(T)) > 0.8$, and categorized as inactive otherwise.

3.3.2 Results

The true activation pattern is displayed in Figure 3.2. The three white blocks in the middle are the areas that are activated in response to the event, while the rest black area is not. After the running of MCMC chains, we are able to plot the posterior activation map on any time point t by assigning value 1 to those voxels with $p(\gamma_{t,\nu}|\boldsymbol{Y}(T)) > 0.8$ and value 0 otherwise. We expect a similar posterior activation map as 3.3 during an activation session (when the task is being performed) and a

total black activation map (with values 0) during an inactivation one.

Figure 3.3 reports one example of the posterior activation map during one activation session (from $t = 44$ to $t = 52$) and Figure 3.4 during an inactivation one (from $t = 54$ to $t = 62$). Our method does a good job at detecting the active voxels. In particular, a small number of active voxels are falsely identified as inactive, and all inactive voxels are correctly identified. During the inactivation session, only very few number of inactive voxels are falsely identified as active.

$\alpha_{t,\nu} = 1, \beta_{t,\nu} = 1$			
d	-3	-3.5	-4
Sensitivity	0.8518	0.8296	0.8141
Specificity	0.988	0.9892	0.9898
$\alpha_{t,\nu} = t, \beta_{t,\nu} = 1$			
d	-3	-3.5	-4
Sensitivity	0.6948	0.66	0.6156
Specificity	0.9894	0.9928	0.995

Table 3.1: Sensitivity analysis on the parameter $\boldsymbol{\alpha}$ of Beta-GOS and d of the MRF

We also perform a sensitivity analysis to find out how different prior and parameter specifications affect the results. Table 3.1 show the results for parameter $\alpha_{t,\nu}$ of Beta-GOS process and d of MRF. In the table, the sensitivity and specificity are reported. Sensitivity (also called the true positive rate) measures the proportion of actual positives which are correctly identified as such, and specificity (sometimes called the true negative rate) measures the proportion of negatives which are correctly identified as such. It seems that the results are more sensitive to the parameter of Beta-GOS process than to the parameter d of MRF. In fact, if we set $\alpha_{t,\nu} = 1, \beta_{t,\nu} = 1$, it is a strong assumption in Beta-GOS process that the expected number of clusters of observations is 2, which in our case is also appropriate because the state consists of both active and inactive. As we have discussed earlier, the parameter d controls the sparsity of the model, with higher values encouraging the selection of voxels with neighbors already selected as active. More specifically, larger values of d correspond to higher sensitivity value, at the cost of a slightly lower specificity.

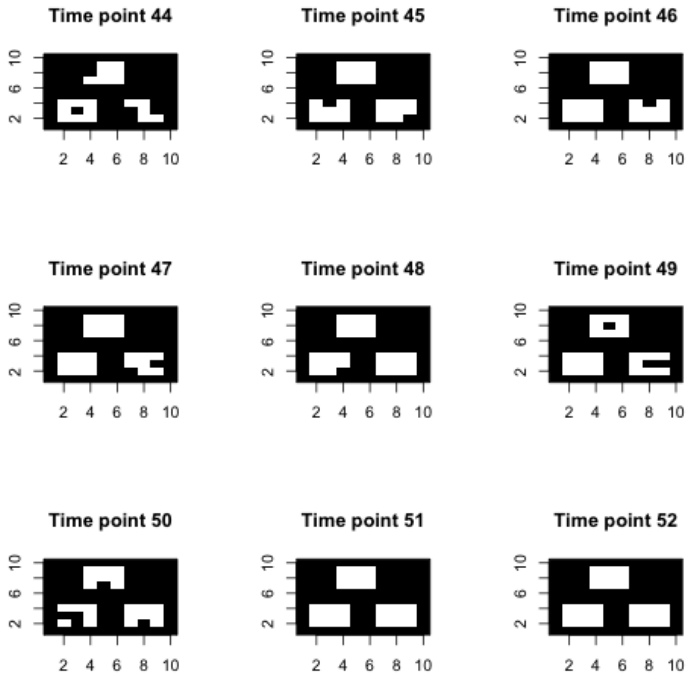

Figure 3.3: Posterior activation maps from $t = 44$ to $t = 52$.

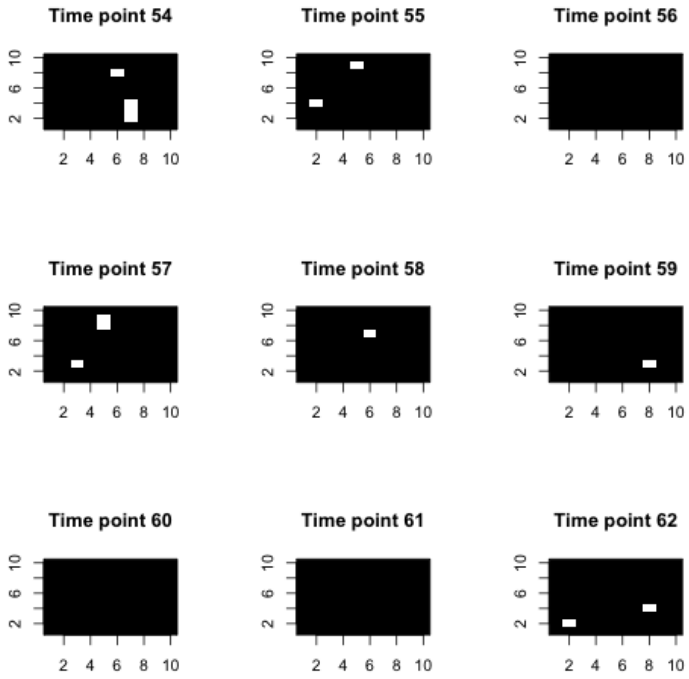

Figure 3.4: Posterior activation maps from $t = 54$ to $t = 62$

3.4 Conclusions

In this chapter, we propose a novel Bayesian nonparametric approach for modeling brain connectivity. In particular, we introduce a mixture of Beta-GOS process and Dirac distribution centered at zero, with the weights depending on how many neighboring voxels are activated. In this way, the model is capable of capturing both temporal and spatial dependence. We apply this model to simulated data set and obtain promising results. Specifically, it well detects the activate and inactivate states.

Chapter 4

A Bootstrap Likelihood approach to Bayesian Computation

Davison et al. (1992) introduced the bootstrap likelihood which combines the nested bootstrap calculation with kernel smoothing methods to calculate estimates of the density of a given statistic for a range of parameter values. These density estimates are used to generate values of an analogue of a likelihood function by curve-fitting methods (see also Efron and Tibshirani (1994), and Davison and Hinkley (1997)). Assume that $\hat{\theta}$ is an estimator of a parameter of interest θ and we seek an approximate likelihood function for θ. The goal is to estimate the sampling density $p\left(\hat{\theta}|\theta\right)$, namely, the the sampling distribution of $\hat{\theta}$ when the true parameter is θ. The basic method can be summarized as follows:

- Suppose θ is the parameter of interest and $\hat{\theta}$ is the parameter estimated by its sample analogue. Generate K bootstrap samples of size n (same size as the original data) to obtain a series of populations $P_1^*, \ldots, P_K^*$ giving bootstrap replications $\hat{\theta}_1^*, \ldots, \hat{\theta}_K^*$ (first-level bootstrap). Any estimation method can be used apart from likelihood estimators.

- For each of the i-th bootstrap samples P_i^* we generate L samples of size n (where L is preferably 1000, as suggested in Davison et al. (1992)). For each sample calculate the analogue of θ, denoted by $\hat{\theta}_{ij}^{**}$ (second-level bootstrap) giving the second stage bootstrap replicates. We form kernel density estimates at each point $\hat{\theta}_i^*$

$$p\left(t|\hat{\theta}_i^*\right) = \frac{1}{L \cdot s} \sum_{j=1}^{L} \ker\left(\frac{t - \hat{\theta}_{ij}^{**}}{s}\right)$$

for $i = 1, \ldots, K$. In this case $\ker(\cdot)$ is any kernel function. We then evaluate $\widehat{p}\left(t|\hat{\theta}_i^*\right)$ for $t = \hat{\theta}$. Since the values $\hat{\theta}_{ij}^{**}$ were generated from a distribution governed by parameter value $\hat{\theta}_i^*$ then $\widehat{p}\left(\hat{\theta}|\hat{\theta}_i^*\right)$ provides an estimate of the likelihood of θ for parameter value $\theta = \hat{\theta}_i^*$. Then the K values $l(\theta_i^*) = \log[\widehat{p}\left(\hat{\theta}|\hat{\theta}_i^*\right)]$ are obtained.

- Apply a smooth curve-fitting algorithm, like a a scatterplot smoother to the pairs $\left(\hat{\theta}_i^*, l(\theta_i^*)\right)$ for $i = 1, \ldots, K$, to obtain the whole log bootstrap likelihood curve.

Although the previous scheme is adapted to the case of *i.i.d.* samples, in the case of dependent data, such as regression-type problems, the outlined method also applies (see e.g. the original paper of Davison et al. (1992)). This is a typical situation in dynamic models, see for instance, Example 4.2.1.

The relation between empirical likelihood and bootstrap likelihood is also explored from a theoretical point of view in Davison et al. (1992) and Owen (2001). They point out that the bootstrap likelihood matches the empirical likelihood to first order. Specifically, in the case of an estimator determined by a monotonic estimating function, standardized so that the leading term is of order one, it is shown by applying empirical cumulants (see Davison et al. (1992)) that empirical and bootstrap likelihoods agree to order $n^{-\frac{1}{2}}$ but not to order n^{-1} in the number of observations, n. In this way, results derived for empirical likelihood, such as the analogue of Wilks' theorem, apply also to the bootstrap likelihood (see Davison et al. (1992)).

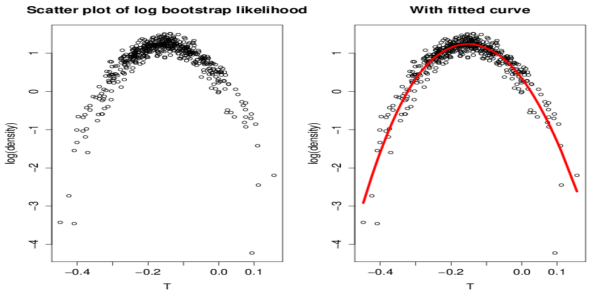

Figure 4.1: The left figure is plotted after the first two steps in the above summarized method. Basically, the first-level bootstrap is for generating the x-axis values and the second-level bootstrap is for the estimation of the density of T at these corresponding x-axis values. The right figure displays the estimated bootstrap likelihood curve.

Figure 4.1 is an illustration of the bootstrap likelihood construction. In the next section, we will use the bootstrap likelihood to develop an algorithm to address Bayesian inference in the spirit of Mengersen, Pudlo and Robert (2013) (see Section 1.4.1).

4.1 Bayesian computation via bootstrap likelihood

Let $BL(\theta_i|\boldsymbol{y})$ denote the estimation of the bootstrap likelihood in the point θ_i given the observed data $\boldsymbol{y}$. Our sampler works as follows

---

Algorithm 2 *Bayesian Computation with bootstrap likelihood*

Estimate the bootstrap likelihood curves of parameters with the samples described in the previous section.
for $i = 1$ *to* M *do*

1. *Generate* θ_i *from the prior distribution* $\pi(\cdot)$
2. *Set the weight* $w_i = BL(\theta_i|\boldsymbol{y})$

end for

The output is a sample of size M of parameters with associated weights, which operate as an importance sampling output. This means that a posterior sample of simulated parameters of size N is sampled with replacement from the M parameters with corresponding weights w_i's. The bootstrap likelihood approach allows us to define an algorithm with the same structure of the one defined in Mengersen, Pudlo and Robert (2013). In contrast with the empirical likelihood method, the bootstrap likelihood doesn't require any set of subjective constraints by virtue of the bootstrap likelihood methodology. This makes the algorithm an automatic and reliable procedure where only a few trivial parameters need to be specified.
Another benefit of using the bootstrap likelihood instead of the empirical likelihood is that the construction of bootstrap likelihood does not depend on the priors. Once the bootstrap likelihood curve is fitted (last step of constructing the bootstrap likelihood), the weight w_i in BC_{bl} sampler is obtained directly by taking values on the fitted curve. In contrast, the BC_{el} sampler requires solving an optimization problem

at each iteration. This leads to significant gain in the computing time when different priors are compared. Anyway, we have to point out that the same approach can be also realized with the empirical likelihood setting when a very large collection of likelihood values has been gathered.

As a toy illustration of the method, we apply the BC_{bl} algorithm to a normal distribution with known variance (equal to 1). Clearly, the parameter of interest is μ and we can see in Figure 4.2 the fitting of the posterior distribution. In this experiment, the computing time of BC_{bl} algorithm is much less than BC_{el} method. The main reason is that the estimation of μ (sample mean in this case) is explicit and straightforward, without need for numerical estimation algorithms.

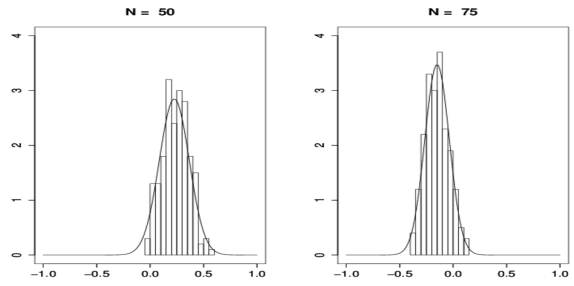

Figure 4.2: Comparison of the true posterior on the normal mean (solid lines) with the empirical distribution of weighted simulations resulting from BC_{bl} algorithm. The normal sample sizes are 50 and 75 respectively, the number of simulated θ's is 200.

In the next Section, the performance of the bootstrap likelihood approach is explored in several examples. In particular, we will see how to manage the parameter estimation in the nested bootstrap. As we will see, this step of the methodology can vary with the problem at the hand.

4.2 Numerical Illustration

4.2.1 Dynamic Models

As mentioned in Section 3, one way to deal with the dependence in dynamic models, is through the application of the bootstrap procedure to the unobserved i.i.d. residuals. For example, we test the GARCH(1,1) model:

$$y_t = \sigma_t \epsilon_t, \quad \epsilon_t \sim N(0,1), \quad \sigma_t^2 = \alpha_0 + \alpha_1 y_{t-1}^2 + \beta_1 \sigma_{t-1}^2$$

under the constraints $\alpha_0, \alpha_1, \beta_1 > 0$ and $\alpha_1 + \beta_1 < 1$ (see Bollerslev (1986)).

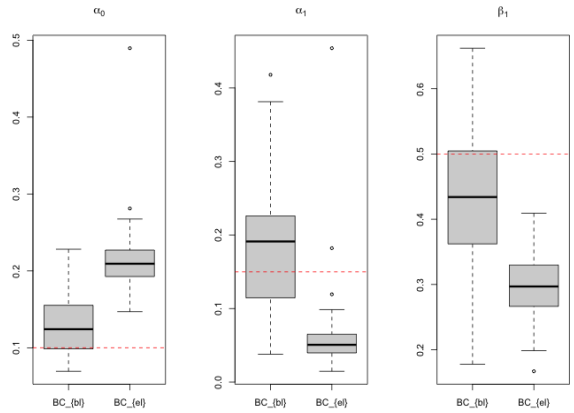

Figure 4.3: Comparison of evaluations of posterior expectations. (with true values in dashed lines) of the parameters $(\alpha_0, \alpha_1, \beta_1)$ of the $GARCH(1,1)$ model with 300 observations.

An exponential $Exp(1)$ and a Dirichlet $Dirich(1,1,1)$ prior distributions are assumed, respectively, on α_0 and $(\alpha_1, \beta_1, 1-\alpha_1-\beta_1)$. In order to compare with BC_{el}, we set the constraints for the empirical likelihood as Mengersen, Pudlo and Robert (2013). The respective number of first and second level of bootstrap replicates are $K = 100$ and $L = 1000$. For each bootstrap replicate, the `R` function `garch` from `tseries` package is used for the estimation of the parameters. This package uses a Quasi-Newton optimizer to find the maximum likelihood estimates of the conditionally normal model. The `garch` function provide a fast estimation of the parameters but it does not always converge consistently. Another alternative may be using the `garchFit` function from `fGarch` package that is slower but converges better.

True values	BC_{bl}	BC_{el}
$\alpha_0 = 0.1$	0.12886(0.00237)	0.19782(0.01039)
$\alpha_1 = 0.15$	0.15307(0.00296)	0.06277(0.01097)
$\beta_1 = 0.5$	0.42874(0.02317)	0.31218(0.03731)

Table 4.1: Summaries of the estimates from two approaches. The results are based on 50 simulated datasets, and displayed with true values in the first column, posterior means from BC_{bl} in the second and posterior means from BC_{el} in the last (with MSE reported inside brackets).

Despite the lack of stability of the `garch` function, in Figure 4.3 we can see that the BC_{bl} algorithm is performing better than the BC_{el} algorithm in terms of the ability to find the correct range of α_0, α_1 and β_1. Furthermore, Table 4.1 illustrates that all parameters are accurately estimated with BC_{bl} with small mean square errors

(MSE), while the estimations with BC_{el} are poorer in this case. One potential reason for the poor performance of BC_{el} is the choice of the score constraints for the empirical likelihood adopted by Mengersen, Pudlo and Robert (2013), which might not guarantee its convergence. Finally, from the computational point of view, we note in our experiments that our approach is faster than the empirical likelihood one. This is not surprising mainly because the bootstrap likelihood procedure depends heavily on the parameter estimation methodology. In this example, the R function `garch` provides a quick estimation of the model parameters and consequently a shorter computational time.

4.2.2 Stochastic differential equations

Stochastic differential equations can be used to model random evolution processes along continuous time, e.g. they are commonly used in many applied areas such as financial models, population dynamics or pharmacokinetics studies. Statistical inference for stochastic differential equations has been undertaken usually from a frequentist point of view, although new Bayesian methodologies have been recently proposed (see Picchini (2014)).

In this section we focus on an example taken from Brouste et al. (2014) and we compare the BC_{bl} procedure with a standard ABC method. We consider the model

$$dX_t = (2 - \theta_2 X_t)dt + \left(1 + X_t^2\right)^{\theta_1} dW_t,$$

where $X_0 = 1$, and we simulate a set of 750 data points assuming $\theta_1 = 0.2$ and $\theta_2 = 0.3$.

We apply first a pure rejection sampling scheme for ABC, with uniform $U(0, 1)$ prior distributions for θ_1 and θ_2 by using the library `EasyABC` for computing tasks.

With regard to BC_{bl}, we use a parametric bootstrap version where the respective number of the first and second levels of bootstrap replicates are $K = 100$ and $L = 200$. For each bootstrap replicate we estimate the parameters by means of a quasi maximum likelihood procedure; in this case, we use the function `qmle` from the `R` package `yuima`.

In Table 4.2 results for both procedures are obtained: estimates by ABC and BC_{bl} are shown with the corresponding *MSE* based in 50 replicates of the model. Here, the estimation of parameters with the ABC approach seems to behave less accurately than BC_{bl}, although we have used quite restricted prior distributions to perform the

True values	ABC	BC_{bl}
$\theta_1 = 0.2$	0.28644 (0.01300)	0.20144 (0.00008)
$\theta_2 = 0.3$	0.41261 (0.02420)	0.34773 (0.02360)

Table 4.2: Summaries of the estimates from two approaches. The results are based on 50 simulated datasets, and displayed with true values in the first column, posterior means from ABC in the second and posterior means from BC_{bl} in the last (with MSE reported inside brackets).

ABC simulations. By using less informative prior distributions, results are still less favourable in the case of the ABC method. Computing times are similar in both cases, although it expands dramatically in the case of ABC methods, when more iterations are required for better approximation of the estimates.

4.2.3 Population Genetics

ABC methods are very popular in population genetics, see e.g. Cornuet et al. (2014). Mengersen, Pudlo and Robert (2013) compare the performance of the BC_{el} sampler with a traditional ABC in the context of evolutionary history of species. They showed that the results are in favor of BC_{el} both in efficiency and effectiveness.

In this section we focus on the study of the distribution of microsatellites, which are repeating sequences of short base pairs of DNA. They are used as molecular markers for kinship and fingerprinting. For a given genetic locus we can consider different types of alleles (genes), namely, alternative forms of the same genetic locus.

The main caution when applying bootstrap likelihood in such setting is the choice of parameter estimates inside each bootstrap level. The true likelihood is intractable in most population genetic settings due to the complexity of the models. However, composite likelihoods have been proved consistent for estimating some parameters such as recombination rates. We will adopt the maximum composite likelihood estimators as parameter estimates in bootstrap likelihood.

Specifically, the intra-locus likelihood is approximated by a product over all pair of genes in the sample at a given locus. Let y_i^k denote the i-th gene at the k-th locus and $\phi = (\tau, \theta)$ the vector of parameters; then the pairwise likelihood of the data at the k-th locus, namely y^k, is defined by

$$l_2\left(y^k|\phi\right) = \prod_{i<j} l_2\left(y_i^k, y_j^k|\phi\right)$$

where

$$l_2\left(y_i^k, y_j^k|\phi\right) = \begin{cases} \frac{\rho(\theta)^{|y_j^k - y_i^k|}}{\sqrt{1+2\theta}} & \text{if same deme} \\ \frac{e^{-\tau\theta}}{\sqrt{1+2\theta}} \sum_{m=-\infty}^{\infty} \rho(\theta)^{|m|} I_{|y_i^k - y_j^k|-m}(\tau\theta) & \text{if different deme} \end{cases}$$

and

$$\rho(\theta) = \frac{\theta}{1+\theta+\sqrt{1+2\theta}}.$$

Note that the expression of the different *deme* case involves an infinite sum, but in practice only the first few terms are required for an accurate approximation, because the value of m corresponds to the number of pairs of mutations in opposite directions, which is usually very small (see Wilson and Balding (1998)).
We compare our proposal with BC_{el} in the first evolutionary scenario studied in Mengersen, Pudlo and Robert (2013), see Figure 4.4.

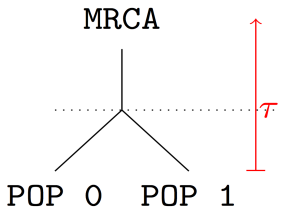

Figure 4.4: Evolutionary scenario of genetic experiment.

Briefly, the genealogy at a given locus is simulated until the most recent common ancestor according to coalescence theory. Then a single mutation event is put at random on one branch of the genealogy. In this scenario, there are two parameters of interest τ and θ. Specifically, τ is the time at which the two populations diverged in the past and $\theta/2$ is the mutation rate of the mutations at a given locus. The simulated datasets are made of ten diploid individuals per population genotyped at fifty independent loci. We use the `DIYABC` software (see Cornuet et al. (2014)) for simulations of the population.

	BC_{bl}	BC_{el}
$\theta = 10$	9.74168(3.76261)	9.38650(3.35539)
$\tau = 0.5$	0.42101(0.02918)	0.54501(0.13742)

Table 4.3: Summaries of the estimates from two approaches. The results are based on 20 simulated datasets, and displayed with true values in the first column, posterior means from BC_{bl} in the second and posterior means from BC_{el} in the last (with MSE reported inside brackets).

All details about implementations of the BC_{el} procedure can be fully found in Mengersen, Pudlo and Robert (2013). By comparing the posterior means and MSE in Table 4.3, one can find a similar accuracy and precision of the estimates from both BC_{bl} and BC_{el} samplers. We then compare the marginal posterior distributions of the parameters θ and τ obtained with both samplers.

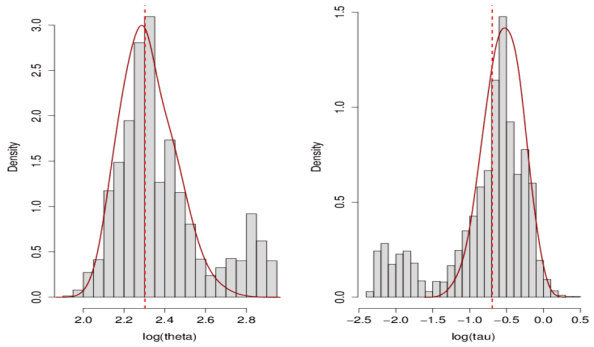

Figure 4.5: Comparison of the marginal distributions obtained by the BC_{el} and the BC_{bl} sampler. The histogram is sampled using BC_{el} and the curve is the result of BC_{bl}.

Figure 4.5 suggests that BC_{el} has difficulties eliminating the tails of both posterior distributions and BC_{bl} is more accurate in terms of the shape. Mengersen, Pudlo and Robert (2013) further suggest the incorporation of empirical likelihood in the adaptive multiple importance sampling (AMIS) to speed up the computation. The bootstrap likelihood could also be incorporated in the same way. However, Figure 4.6 shows that AMIS improves substantially the results computed with the basic BC_{el} sampler, but not so much with respect to the BC_{bl} sampler. For instance, in the case of parameter τ, using AMIS does not improve the performance of BC_{bl} with respect to the true value of the parameter. It appears that the basic BC_{bl} sampler is enough capable of building a reasonable posterior, which suggests that it is unnecessary to introduce the AMIS in the bootstrap likelihood setting.
About the computing time, in general, the speed of BC_{bl} depends on many factors, mainly including the numbers of first and second level bootstrap replicates, and the difficulty to estimate the parameter inside each bootstrap level. In this example, the **R** function **optim** is employed to estimate the maximum composite likelihood estimator. The speed of BC_{el} depends on the difficulty to optimize under the constraints and the size of the Monte Carlo sample. For this reason we resort to the **R** library **emplik** for the calculation of the empirical likelihood. In this experiment, we also noticed that the computing time of BC_{bl} is more or less twice the time needed

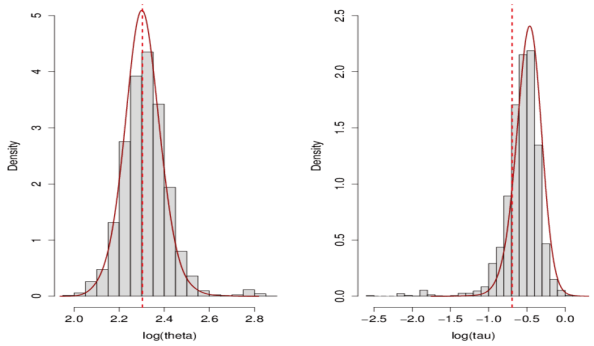

Figure 4.6: Comparison of the $BC_{el} - AMIS$ and the $BC_{bl} - AMIS$ sampler. The histogram is sampled using $BC_{el} - AMIS$ and the curve is the result of $BC_{bl} - AMIS$.

for BC_{el} under our parameter setting (50 bootstrap replicates in the first level and 100 replicates in the second for bootstrap likelihood, 30000 Monte Carlo samples in BC_{el}).

4.2.4 Ising and Potts Model

Ising and Potts models are discrete Gibbs random field models with a statistical physics origin, which are now widely used for applications in spatial modelling, image processing, computational biology, and computational neuroscience. Consider the simple case of a random field where the pixels of the image $\boldsymbol{x}$ can only take two colours (white and black, say). Let $\{\boldsymbol{x} = x_{ij} : (i,j) \in D\}$ denote the observed binary data, where x_{ij} is a pixel and D is an $M \times N$ lattice indexing the pixels. The conditional distribution of a pixel is then Bernoulli, with the parameter being a function of the number of neighbouring pixels that have the same value. It is defined as

$$f(x_{ij} = k | x_{n(i,j)}) \propto \exp(\beta n_{i,j}^k), \qquad \beta > 0, \quad k = 0, 1$$

where

$$n_{i,j}^k = \sum_{l \in n(i,j)} \mathbb{I}_{x_l = k}$$

is the number of neighbours of x_{ij} with colour k and $n(i,j) = \{(i+1,j),(i-1,j),(i,j+1),(i,j-1)\}$ is the defined neighbourhood structure. In Statistical Mechanics, β is a strictly positive parameter which can be interpreted as the inverse

of the temperature. The Ising model is defined through these full conditionals

$$f(x_{ij} = 1|x_{n(i,j)}) = \frac{\exp(\beta n^1_{i,j})}{\exp(\beta n^0_{i,j}) + \exp(\beta n^1_{i,j})}$$

and the joint distribution therefore satisfies

$$f(\boldsymbol{x}) \propto \exp\left(\beta \sum_{(i,j)\sim(i',j')} \mathbb{I}_{\{x_{ij}=x_{i'j'}\}}\right)$$

where the summation is taken over all the neighbour pairs, namely, a neighbourhood relation on pixels is denoted as $\sim$, where $i \sim j$ denotes that i and j are *neighbours*. This joint distribution can be obtained from the conditional distributions, by applying the Hammersley-Clifford representation (see Grimmett (2010)). The Potts model is the natural extension of the Ising Model where more than two colours are considered, see Marin and Robert (2014).

The normalizing constant $Z(\beta)$ of the above distribution depends on β and it is numerically tractable only for very small lattices D, which becomes a major obstacle when making inference on β. The maximum pseudo-likelihood estimator (MPLE) Besag (1977) provides a way to handle the problem. MPLE takes the value that maximizes the pseudo-likelihood function

$$L(\beta|\boldsymbol{x}) = \prod_{i=1}^{M}\prod_{j=1}^{N} f(x_{ij} = 1|x_{n(i,j)}, \beta)$$

We will adopt MPLE as the estimation tool to construct the bootstrap likelihood for β later. Marin and Robert (2014) introduce ABC as a way to simulate the posterior distribution β. However, simulating a data set is unfortunately non-trivial for Markov random fields, as it usually requires a certain number of steps of an MCMC sampler.

We compare the performance of ABC and BC_{bl} in a simulation dataset of size 25×25 where the true parameter β is set as 0.5. The simulation is done using the Gibbs sampler, starting with a random configuration with each pixel being drawn independently from $\{0, 1\}$, and then iterating for 200 Gibbs cycles. The sufficient statistic S is

$$S(\boldsymbol{x}) = \sum_{(i,j)\sim(i',j')} \mathbb{I}_{\{x_{ij}=x_{i'j'}\}}.$$

In order to preserve the spatial structure of data we consider blocks of pixels as bootstrap sampling units, and we apply moving block bootstrap (MBB) methods as suggested by Lahiri (2003). A simulation study about optimum block dimensions can be found in Zhu and Morgan (2004).
Then, in the simulated data, as the corresponding structure is a square grid of pixels, we use a square moving window of length side equal to 5, as it renders good performance to estimate the original parameters. By other hand, we use the MPLE technique for estimating the parameters in each iteration. The numbers of bootstrap replicates for the 1st level and 2nd level bootstrap are 100 and 200, respectively. A $U(0,2)$ is assumed for the parameter β. The choice of the interval $[0,2]$ is motivated by the critical value $\beta = 1$ that represents the phase transition of the Ising Model. Figure 4.7 shows that the estimation carried with BC_{bl} and ABC algorithms provides similar results. It is worth to mention that the BC_{bl} has a computational cost which is less than ABC since the Gibbs sampling for the Ising model has a cost which increases quadratically as the lattice grows. The same problem arises with Potts model where more than two colours are considered.

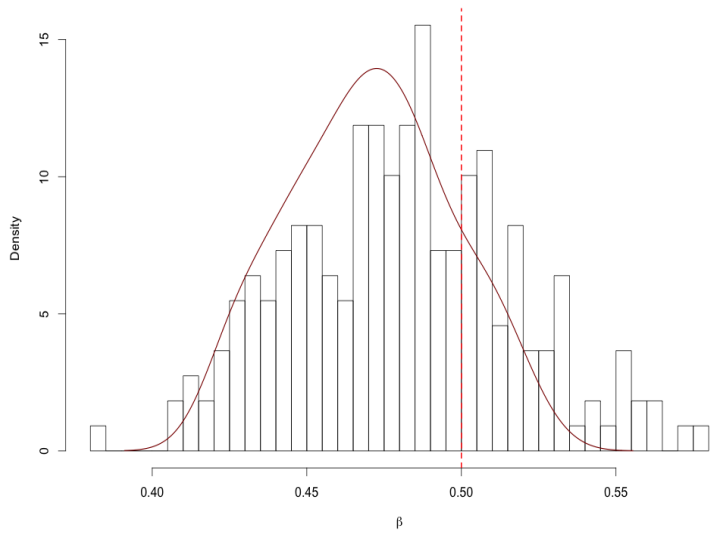

Figure 4.7: Comparison of the BC_{bl} (curve) with the histogram of the simulations from ABC algorithm with 10^4 iterations and a 1% quantile on the difference between the sufficient statistics as its tolerance bound ϵ, based on the uniform prior $U(0,2)$.

We conclude this section with a real data example. In particular, we consider a set of soil phosphate measurements collected during the Laconia Archaeological Survey

in Greece (year 1987). A complete description of data can be found i.e. in Buck et al. (1988). This dataset has been analysed by using different techniques, for instance Buck et al. (1988) carried a Bayesian change-point analysis to describe the dataset. On the other hand, Besag et al. (1991) adopted a Bayesian image analysis approach. Recently, McGrory et al. (2009) studied the dataset with variational Bayes methods.

In this application, we use the moving block bootstrap and MPLE techniques in a similar way as in the simulation study. The window length of the moving block is set as 8. The numbers of bootstrap replicates for the 1st level and 2nd level bootstrap are 100 and 100, respectively. The distribution of values of β is shown in Figure 4.8.

The distribution of parameter β is roughly located between 0.40 and 0.55; it may be noted that results are quite similar to those obtained inMcGrory et al. (2009) who use a variational Bayes method. In their case, the estimation of parameter β, by using variational Bayes and MCMC methods, also renders similar estimates between 0.44 and 0.59 (see Table 3 of McGrory et al. (2009)).

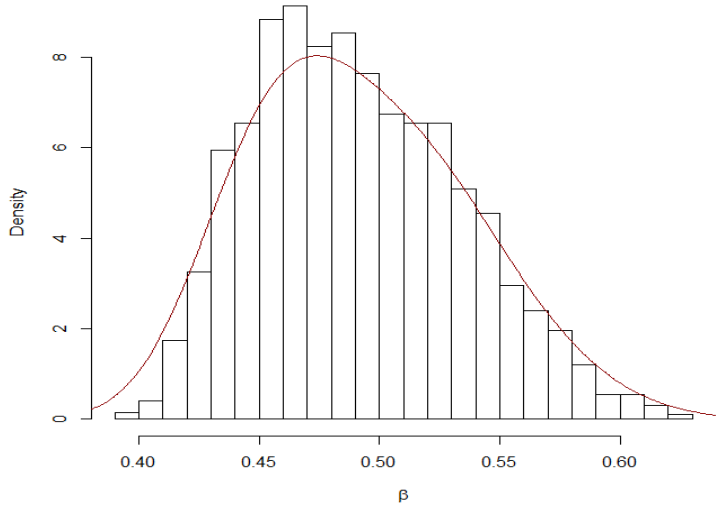

Figure 4.8: Histogram and density estimation of parameter β after applying BC_{bl} in the Laconia Archaeological data.

4.3 Conclusions and future research

In this chapter, we introduced a bootstrap likelihood approach to address inference in a Bayesian setting. The sampling scheme has a structure which is similar to the recent work of Mengersen, Pudlo and Robert (2013). These type of algorithms can be used as an alternative to the standard ABC methods when difficulties arise in setting the parameters (distance, summary statistics and tolerance level). In particular, the empirical likelihood approach of Mengersen, Pudlo and Robert (2013) allows to avoid this problem but, on the other hand, requires to choose a set of constraints. Different choices could sensibly affect the inference and, consequently, the parameter estimation. The main advantage of the bootstrap likelihood approach is that it is an automatic procedure that does not require the careful choice of the constraints. In this chapter, merits and problems of the new approach are discussed through simulation experiments. In particular, the method is tested on dynamic models, a stochastic differential equations and a population genetics problem. Furthermore, in a random field context, an application to real data is provided.

Conclusions and further research

In this thesis, firstly we extended the bivariate vector of Leisen and Lijoi (2011) to the multivariate setting. We provided the derivation of the Laplace transform which is non-trivial in the multivariate setting and furthermore, an expression of the Exchangeable Partition Probability Function (EPPF). Also, a new MCMC algorithm has been introduced for evaluating the EPPF, which provides a useful tool to further study the clustering behavior of the new prior. Secondly, we proposed a novel Bayesian nonparametric approach for modeling brain connectivity. In particular, we introduced a mixture of Beta-GOS process and a Dirac distribution centered at zero, where the weights are modeled with a random field prior. In this way, the model is capable of capturing both temporal and spatial dependence. We applied the model to simulated data set and obtained promising results. Since it is still an ongoing work, the next step is to apply the method to the real data. Finally, we introduced a bootstrap likelihood approach to address inference in a Bayesian setting. Our method can be used as an alternative to the standard ABC methods when difficulties arise in setting the parameters (distance, summary statistics and tolerance level), or as an alternative to the sampler based on empirical likelihood when the choice of constrains is not clear. The merits and problems of the new approach were discussed through simulation experiments. Also, we provided an application to real data.

As for the further research, we will try to develop a MCMC algorithm to sample from the vectors that we proposed. Besides, a new class of dependent random measures which is called compound random measures (see Griffin and Leisen (2014)) suggests another way to construct new priors. We can develop new models from that point of view. Another further work direction is on the BC_{bl} sampler. The current methodology requires a choice of estimate of the parameters, such as the maximum likelihood estimator or pseudo likelihood based estimator. We hope to modify the sampler so that its application can be extended to the models where the estimation

of the parameters can only be done in nonparametric ways. This will greatly increase the area of the application of BC_{bl}.

Appendix A

Proofs of Chapter 2

Lemma 1 *Let $(\tilde{\mu}_1, ..., \tilde{\mu}_d)$ be a vector of CRMs with Laplace exponent $\psi_{\rho,d}(\boldsymbol{\lambda})$. If $C_\epsilon \in \mathscr{X}$ is such that diam$(C_\epsilon)\downarrow 0$ as $\epsilon \downarrow 0$, then*

$$\mathbb{E}\left[e^{-\lambda_1\tilde{\mu}_1(C_\epsilon)-\cdots-\lambda_d\tilde{\mu}_d(C_\epsilon)}\prod_{i=1}^{d}\{\tilde{\mu}_i(C_\epsilon)\}^{q_i}\right] =$$
$$(-1)^{q_1+...+q_d-1}\alpha(C_\epsilon)e^{-\alpha(C_\epsilon)\psi_{\rho,d}(\boldsymbol{\lambda})} \times \frac{\partial^{q_1+...+q_d}}{\partial\lambda_1^{q_1}\cdots\partial\lambda_d^{q_d}}\psi_{\rho,d}(\boldsymbol{\lambda}) + o(\alpha(C_\epsilon))$$

as $\epsilon \downarrow 0$.

Proof. The proof follows from a simple application of a multivariate version of the Faá di Bruno formula, see Constantine and Savits (1996).

$$\mathbb{E}\left[e^{-\lambda_1\tilde{\mu}_1(C_\epsilon)-\cdots-\lambda_d\tilde{\mu}_d(C_\epsilon)}\prod_{i=1}^{d}\{\tilde{\mu}_i(C_\epsilon)\}^{q_i}\right] = (-1)^{q_1+...+q_d}\frac{\partial^{q_1+...+q_d}}{\partial\lambda_1^{q_1}\cdots\partial\lambda_d^{q_d}}e^{-\alpha(C_\epsilon)\psi_{\rho,d}(\boldsymbol{\lambda})}$$

The right-hand side above coincides with

$$e^{-\alpha(C_\epsilon)\psi_{\rho,d}(\boldsymbol{\lambda})}q_1!\cdots q_d!\sum_{k=1}^{q_1+...+q_d}(-1)^k[\alpha(C_\epsilon)]^k\times$$
$$\sum_{j=1}^{q_1+...+q_d}\sum_{p_j(q_1,...,q_d,k)}\prod_{i=1}^{j}\frac{1}{\beta_i!(s_{1,i}!\cdots s_{d,i}!)^{\beta_i}}\left(\frac{\partial^{s_{1,i}+...+s_{d,i}}}{\partial\lambda_1^{s_{1,i}}\cdots\partial\lambda_d^{s_{d,i}}}\psi_{\rho,d}(\boldsymbol{\lambda})\right)^{\beta_i}$$

where $p_j(q_1, ..., q_d, k)$ is the set of vectors $(\boldsymbol{\beta}, \boldsymbol{s_1}, ..., \boldsymbol{s_j})$ with $\boldsymbol{\beta}= (\beta_1, ..., \beta_j)$ a vector whose positive coordinates are such that $\sum_{i=1}^{j}\beta_i = k$ and the $\boldsymbol{s_i} = (s_{1,i}, ..., s_{d,i})$ are vectors such that $\mathbf{0} \prec \boldsymbol{s_1} \prec \cdots \prec \boldsymbol{s_j}$. Obviously, in the previous sum, all terms

with $k \geq 2$ are $o(\alpha(C_\epsilon))$ as $\epsilon \downarrow 0$. Furthermore, if we suppose that the Lévy measure is of finite variation, i.e.
$\int_{\|\boldsymbol{y}\|\leq 1} \|\boldsymbol{y}\|\rho_d(y_1,...,y_d)dy_1\cdots dy_d < \infty$ where $\|\boldsymbol{y}\|$ stands for the Euclidean norm of the vector $\boldsymbol{y} = (y_1,...,y_d)$, then one also has
$\int_{\|\boldsymbol{y}\|\leq 1} y_1^{n_1}\cdots y_d^{n_d}\rho_d(y_1,...,y_d)dy_1\cdots dy_d < \infty$ for any $n_i,\ i = 1,...,d$ positive integers.

$$\mathbb{E}\left[e^{-\lambda_1\tilde{\mu}_1(C_\epsilon)-\cdots-\lambda_d\tilde{\mu}_d(C_\epsilon)}\prod_{i=1}^{d}\{\tilde{\mu}_i(C_\epsilon)\}^{q_i}\right] = \alpha(C_\epsilon)e^{-\alpha(C_\epsilon)\psi_{\rho,d}(\boldsymbol{\lambda})}g_\rho(q_1,...,q_d;\boldsymbol{\lambda}) + o(\alpha(C_\epsilon))$$

as $\epsilon \downarrow 0$, for any $\lambda_i > 0$, $i = 1,...,d$, where

$$g_\rho(q_1,...,q_d;\boldsymbol{\lambda}) = \int_{(0,\infty)^d} y_1^{q_1}\cdots y_d^{q_d}e^{-\lambda_1 y_1-\cdots-\lambda_d y_d}\rho_d(y_1,...,y_d)dy_1\cdots dy_d$$

□

Theorem 4 *Let $g_\rho(q_1,...,q_d;\boldsymbol{\lambda})$ be defined as (2.14). Let $I \in \{1,\ldots,d\}$ be such that $\lambda_I = \max(\lambda_1,\ldots,\lambda_d)$. Hence,*

$$g_\rho(q_1,...,q_d;\boldsymbol{\lambda}) = \frac{(\sigma)_d}{\Gamma(1-\sigma)}\frac{\Gamma(|\boldsymbol{q}|-\sigma)}{\lambda_I^{|\boldsymbol{q}|-\sigma}}\frac{\prod_{i=1}^{d}\Gamma(q_i+1)}{\Gamma(|\boldsymbol{q}|+d)} \times F_D(|\boldsymbol{q}|-\sigma;\boldsymbol{q}_{-I}+\mathbf{1};|\boldsymbol{q}|+d;\mathbf{1}-\frac{\boldsymbol{\lambda_{-I}}}{\lambda_I})$$

where $\boldsymbol{q}_{-I}+\mathbf{1}$ and $\mathbf{1}-\frac{\boldsymbol{\lambda_{-I}}}{\lambda_I}$ are the vectors of parameters

$$\boldsymbol{q}_{-I}+\mathbf{1} = (q_1+1,\ldots,q_{I-1}+1,q_{I+1}+1,\ldots,q_d+1)$$
$$\mathbf{1}-\frac{\boldsymbol{\lambda_{-I}}}{\lambda_I} = \left(1-\frac{\lambda_1}{\lambda_I},...,1-\frac{\lambda_{I-1}}{\lambda_I},1-\frac{\lambda_{I+1}}{\lambda_I},\cdots,1-\frac{\lambda_d}{\lambda_I}\right)$$

Proof. The Lauricella function of fourth kind is defined as

$$F_D(a,b_1,\ldots,b_d,c,z_1,\ldots,z_d) = \sum_{m_1=0}^{\infty}\cdots\sum_{m_d=0}^{\infty}\frac{(a)_{|\boldsymbol{m}|}(b_1)_{m_1}\cdots(b_d)_{m_d}}{(c)_{|\boldsymbol{m}|}m_1!\cdots m_d!}z_1^{m_1}\cdots z_d^{m_d}$$

for $|z_i| < 1$, $i = 1,\ldots,d$ and $|\boldsymbol{m}| = m_1+\cdots+m_d$. An integral representation is also

available for the Lauricella function of fourth kind, that is

$$F_D(a, b_1, \ldots, b_d, c, z_1, \ldots, z_d) = \frac{\Gamma(c)}{\Gamma(c - \sum_{i=1}^d b_i) \prod_{i=1}^d \Gamma(b_i)} \times \int_{\Delta_d} (1 - |\boldsymbol{y}|)^{c-1-\sum_{i=1}^d b_i} \prod_{i=1}^d y_i^{b_i - 1} (1 - < \boldsymbol{y}, \boldsymbol{z} >)^{-a} d\boldsymbol{y}$$

Without loss of generality, suppose that $I = d$. A simple change of variable, namely $z_i = y_i/s$ for $i = 1, ..., d-1$ and $s = |\boldsymbol{y}|$, yields

$$\begin{aligned} g_\rho(q_1, ..., q_d; \boldsymbol{\lambda}) &= \frac{(\sigma)_d}{\Gamma(1-\sigma)} \int_{\Delta_{d-1}} (1 - |\boldsymbol{z}|)^{q_d} \prod_{i=1}^{d-1} z_i^{q_i} \\ &\qquad \times \int_0^{+\infty} s^{|\boldsymbol{q}| - \sigma - 1} e^{-s[<\boldsymbol{\lambda}, \boldsymbol{z}> + \lambda_d(1 - |\boldsymbol{z}|)]} ds d\boldsymbol{z} \\ &= \frac{(\sigma)_d \Gamma(|\boldsymbol{q}| - \sigma)}{\Gamma(1-\sigma)} \int_{\triangle_{d-1}} \frac{(1 - |\boldsymbol{z}|)^{q_d} \prod_{i=1}^{d-1} z_i^{q_i}}{[< \boldsymbol{\lambda}, \boldsymbol{z} > + \lambda_d(1 - |\boldsymbol{z}|)]^{|\boldsymbol{q}| - \sigma}} d\boldsymbol{z} \end{aligned}$$

□

Proof of the Proposition 1. Before proving the statement of the proposition we need some preliminaries. For every $a > 0$ and $d > 1$ such that $d \in \mathbb{N}$, define the integral

$$\Phi_d^a(\boldsymbol{\lambda}) = \int_{\triangle_{d-1}} (a+1)_{d-1} \left[\lambda_1 z_1 + \cdots + \lambda_{d-1} z_{d-1} + \lambda_d (1 - z_1 - \cdots - z_{d-1})\right]^a d\boldsymbol{z} \tag{A.1}$$

where $\boldsymbol{z} = (z_1, \ldots, z_{d-1})$ and $\triangle_{d-1} = \{\boldsymbol{z} \in (0,1)^{d-1} : z_1 + \cdots + z_{d-1} < 1\}$. The Laplace exponent is closely related to this integral, indeed a simple change of variable, namely $z_i = y_i/s$ for $i = 1, ..., d-1$ and $s = |\boldsymbol{y}|$ in equation (2.10), yields

$$\psi_{\rho,d}(\boldsymbol{\lambda}) = \Phi_d^\sigma(\boldsymbol{\lambda})$$

i.e. the Laplace Exponent coincides with the integral (A.1) when $a = \sigma$. Proving the statement of Proposition 1, therefore, is equivalent to prove that

$$\Phi_d^\sigma(\boldsymbol{\lambda}) = \Phi_d^\sigma(\tilde{\boldsymbol{\lambda}}, \boldsymbol{n}) = \left(\prod_{i=1}^l \frac{1}{\Gamma(n_i)} \frac{\partial^{n_i - 1}}{\partial^{n_i - 1} \tilde{\lambda}_i} \right) \left(\phi_l^\sigma(\tilde{\boldsymbol{\lambda}}) \prod_{i=1}^l \tilde{\lambda}_i^{n_i - 1} \right), \tag{A.2}$$

We prove it by induction. The case of $d = 2$ has been proved by Leisen and Lijoi (2011) and it is displayed in Equation (2.11).
Now, suppose that (A.2) holds for d with $l \leq d$ distinct values $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_l$ among the $\lambda_1, \ldots, \lambda_d$ with multiplicities $n_1, \ldots, n_l$. For the case of $d+1$, there are two scenarios for λ_{d+1}

a) $\lambda_{d+1} \in \{\lambda_1, \ldots, \lambda_d\}$

b) $\lambda_{d+1} \notin \{\lambda_1, \ldots, \lambda_d\}$

Case a). In the first case, without loss of generality, we can assume that $\lambda_{d+1} = \lambda_1 = \tilde{\lambda}_1 \neq \tilde{\lambda}_l = \lambda_d$. Hence,

$$
\begin{aligned}
&\Phi^{\sigma}_{d+1}(\tilde{\lambda}_1, ..., \tilde{\lambda}_l; n_1 + 1, ..., n_l) \\
&\quad = \frac{1}{\tilde{\lambda}_l - \tilde{\lambda}_1} \Big[\Phi^{\sigma+1}_d(\tilde{\lambda}_1, ..., \tilde{\lambda}_l; n_1, ..., n_l) - \Phi^{\sigma+1}_d(\tilde{\lambda}_1, ..., \tilde{\lambda}_l; n_1 + 1, n_2, ..., n_{l-1}, n_l - 1) \Big] \\
&\quad = \frac{1}{\tilde{\lambda}_l - \tilde{\lambda}_1} \Bigg\{ \frac{1}{\prod_{i=1}^{l} \Gamma(n_i)} \frac{\partial^{n_1 - 1} \cdots \partial^{n_l - 1}}{\partial^{n_1 - 1} \tilde{\lambda}_1 \cdots \partial^{n_l - 1} \tilde{\lambda}_l} \left[\phi^{\sigma+1}_l(\tilde{\boldsymbol{\lambda}}) \prod_{i=1}^{l} \tilde{\lambda}_i^{n_i - 1} \right] \\
&\qquad - \frac{n_l - 1}{n_1} \frac{1}{\prod_{i=1}^{l} \Gamma(n_i)} \frac{\partial^{n_1} \partial^{n_2 - 1} \cdots \partial^{n_{l-1} - 1} \partial^{n_l - 2}}{\partial^{n_1} \tilde{\lambda}_1 \partial^{n_2 - 1} \tilde{\lambda}_2 \cdots \partial^{n_{l-1} - 1} \tilde{\lambda}_{l-1} \partial^{n_l - 2} \tilde{\lambda}_l} \left[\phi^{\sigma+1}_l(\tilde{\boldsymbol{\lambda}}) \frac{\tilde{\lambda}_1}{\tilde{\lambda}_l} \prod_{i=1}^{l} \tilde{\lambda}_i^{n_i - 1} \right] \Bigg\} \\
&\quad = \frac{1}{\tilde{\lambda}_l - \tilde{\lambda}_1} \Bigg\{ \frac{1}{n_1} \frac{1}{\prod_{i=1}^{l} \Gamma(n_i)} \frac{\partial^{n_1 - 1} \cdots \partial^{n_l - 1}}{\partial^{n_1 - 1} \tilde{\lambda}_1 \cdots \partial^{n_l - 1} \tilde{\lambda}_l} \left[\phi^{\sigma+1}_l(\tilde{\boldsymbol{\lambda}}) \frac{\partial}{\partial \tilde{\lambda}_1} \left(\tilde{\lambda}_1^{n_1} \right) \prod_{i=2}^{l} \tilde{\lambda}_i^{n_i - 1} \right] \\
&\quad - \frac{1}{n_1} \frac{1}{\prod_{i=1}^{l} \Gamma(n_i)} \frac{\partial^{n_1} \partial^{n_2 - 1} \cdots \partial^{n_{l-1} - 1} \partial^{n_l - 2}}{\partial^{n_1} \tilde{\lambda}_1 \partial^{n_2 - 1} \tilde{\lambda}_2 \cdots \partial^{n_{l-1} - 1} \tilde{\lambda}_{l-1} \partial^{n_l - 2} \tilde{\lambda}_l} \left[\phi^{\sigma+1}_l(\tilde{\boldsymbol{\lambda}}) \tilde{\lambda}_1 \prod_{i=1}^{l-1} \tilde{\lambda}_i^{n_i - 1} \frac{\partial}{\partial \tilde{\lambda}_l} \left(\tilde{\lambda}_l^{n_l - 1} \right) \right] \Bigg\}
\end{aligned}
\tag{A.3}
$$

Let $\boldsymbol{\lambda}_{-1} = (\tilde{\lambda}_2, \ldots, \tilde{\lambda}_l)$. A key element for the next computations is the following identity

$$
\phi^{\sigma+1}_l(\tilde{\boldsymbol{\lambda}}) = \tilde{\lambda}_1 \phi^{\sigma}_l(\tilde{\boldsymbol{\lambda}}) + \phi^{\sigma+1}_l(\tilde{\boldsymbol{\lambda}}_{-1}) \tag{A.4}
$$

Notice that $\phi^{\sigma+1}_l(\tilde{\boldsymbol{\lambda}}_{-1})$ does not depend on $\tilde{\lambda}_1$.
The first term of (A.3) could be written as

$$
\begin{aligned}
&\frac{1}{\tilde{\lambda}_l - \tilde{\lambda}_1} \Bigg\{ \frac{1}{n_1} \frac{1}{\prod_{i=1}^{l} \Gamma(n_i)} \frac{\partial^{n_1} \cdots \partial^{n_l - 1}}{\partial^{n_1} \tilde{\lambda}_1 \cdots \partial^{n_l - 1} \tilde{\lambda}_l} \left[\phi^{\sigma+1}_l(\tilde{\boldsymbol{\lambda}}) \tilde{\lambda}_1^{n_1} \prod_{i=2}^{l} \tilde{\lambda}_i^{n_i - 1} \right] \\
&\qquad - \frac{1}{n_1} \frac{1}{\prod_{i=1}^{l} \Gamma(n_i)} \frac{\partial^{n_1 - 1} \cdots \partial^{n_l - 1}}{\partial^{n_1 - 1} \tilde{\lambda}_1 \cdots \partial^{n_l - 1} \tilde{\lambda}_l} \left[\tilde{\lambda}_1^{n_1} \prod_{i=2}^{l} \tilde{\lambda}_i^{n_i - 1} \frac{\partial}{\partial \tilde{\lambda}_1} \phi^{\sigma+1}_l(\tilde{\boldsymbol{\lambda}}) \right] \Bigg\}
\end{aligned}
\tag{A.5}
$$

By applying the general Leibniz rule and noting that the n_1-th derivative of $\tilde{\lambda}_1^{n_1-1}$ is 0, one gets,

$$\begin{aligned}\frac{\partial^{n_1-1}}{\partial^{n_1-1}\tilde{\lambda}_1}\left[\tilde{\lambda}_1^{n_1}\prod_{i=2}^{l}\tilde{\lambda}_i^{n_i-1}\frac{\partial}{\partial\tilde{\lambda}_l}\phi_l^{a+1}(\tilde{\boldsymbol{\lambda}})\right] &= \sum_{k=0}^{n_1-1}\binom{n_1-1}{k}\frac{\partial^k}{\partial^k\tilde{\lambda}_1}(\tilde{\lambda}_1^{n_1})\frac{\partial^{n_1-k}}{\partial^{n_1-k}\tilde{\lambda}_l}\phi_l^{a+1}(\tilde{\boldsymbol{\lambda}}) \\ &= \tilde{\lambda}_1\sum_{k=0}^{n_1}\binom{n_1}{k}\frac{\partial^k}{\partial^k\tilde{\lambda}_1}(\tilde{\lambda}_l^{n_1-1})\frac{\partial^{n_1-k}}{\partial^{n_1-k}\tilde{\lambda}_1}\phi_l^{a+1}(\tilde{\boldsymbol{\lambda}}) \\ &= \tilde{\lambda}_1\frac{\partial^{n_1}}{\partial^{n_1}\tilde{\lambda}_1}\tilde{\lambda}_l^{n_1-1}\phi_l^{a+1}(\tilde{\boldsymbol{\lambda}})\end{aligned}$$

and from identity (A.4) , equation (A.5) could be further written as

$$\begin{aligned}\frac{1}{\tilde{\lambda}_l-\tilde{\lambda}_1}\Bigg\{\frac{1}{n_1}\frac{1}{\prod_{i=1}^{l}\Gamma(n_i)}\frac{\partial^{n_1}\cdots\partial^{n_l-1}}{\partial^{n_1}\tilde{\lambda}_1\cdots\partial^{n_l-1}\tilde{\lambda}_l}\left[\phi_l^{\sigma+1}(\tilde{\boldsymbol{\lambda}}_{-1})\tilde{\lambda}_1^{n_1}\prod_{i=2}^{l}\tilde{\lambda}_i^{n_i-1}\right] \\ -\frac{1}{n_1}\frac{\tilde{\lambda}_1}{\prod_{i=1}^{l}\Gamma(n_i)}\frac{\partial^{n_1}\cdots\partial^{n_l-1}}{\partial^{n_1}\tilde{\lambda}_1\cdots\partial^{n_l-1}\tilde{\lambda}_l}\left[\phi_l^{\sigma}(\tilde{\boldsymbol{\lambda}})\tilde{\lambda}_1\prod_{i=1}^{l}\tilde{\lambda}_i^{n_i-1}\right]\Bigg\}\end{aligned} \tag{A.6}$$

In a similar way, the second term of (A.3) could be written as

$$\begin{aligned}-\frac{1}{\tilde{\lambda}_l-\tilde{\lambda}_1}\Bigg\{\frac{1}{n_1}\frac{1}{\prod_{i=1}^{l}\Gamma(n_i)}\frac{\partial^{n_1}\cdots\partial^{n_l-1}}{\partial^{n_1}\tilde{\lambda}_1\cdots\partial^{n_l-1}\tilde{\lambda}_l}\left[\phi_l^{\sigma+1}(\tilde{\boldsymbol{\lambda}}_{-1})\tilde{\lambda}_1^{n_1}\prod_{i=2}^{l}\tilde{\lambda}_i^{n_i-1}\right] \\ -\frac{1}{n_1}\frac{\tilde{\lambda}_l}{\prod_{i=1}^{l}\Gamma(n_i)}\frac{\partial^{n_1}\cdots\partial^{n_l-1}}{\partial^{n_1}\tilde{\lambda}_1\cdots\partial^{n_l-1}\tilde{\lambda}_l}\left[\phi_l^{\sigma}(\tilde{\boldsymbol{\lambda}})\tilde{\lambda}_1\prod_{i=1}^{l}\tilde{\lambda}_i^{n_i-1}\right]\Bigg\}\end{aligned} \tag{A.7}$$

Combining equation (A.6) with (A.7) we get the thesis of case a), i.e.

$$\Phi_{d+1}^{a}(\tilde{\lambda}_1,...,\tilde{\lambda}_l;n_1+1,...,n_l)=\frac{1}{n_1}\frac{1}{\prod_{i=1}^{l}\Gamma(n_i)}\frac{\partial^{n_1}\cdots\partial^{n_l-1}}{\partial^{n_1}\tilde{\lambda}_1\cdots\partial^{n_l-1}\tilde{\lambda}_l}\left[\phi_l^{a}(\tilde{\boldsymbol{\lambda}})\tilde{\lambda}_1\prod_{i=1}^{l}\tilde{\lambda}_i^{n_i-1}\right]$$

Case b). Without loss of generality, we assume that $\lambda_{d+1}=\tilde{\lambda}_{l+1}\neq\tilde{\lambda}_l=\lambda_d$. Hence,

$$\begin{aligned}&\Phi_{d+1}^{\sigma}(\tilde{\lambda}_1,...,\tilde{\lambda}_l,\tilde{\lambda}_{l+1};n_1,...,n_l,1)= \\ &\frac{1}{\tilde{\lambda}_l-\tilde{\lambda}_{l+1}}\Big[\Phi_d^{\sigma+1}(\tilde{\lambda}_1,...,\tilde{\lambda}_l;n_1,...,n_l)-\Phi_d^{\sigma+1}(\tilde{\lambda}_1,...,\tilde{\lambda}_l,\tilde{\lambda}_{l+1};n_1,...,n_l-1,1)\Big]\end{aligned}$$

By working in a similar fashion of "case a)", one gets the thesis for "case b)", i.e.

$$\Phi^{\sigma}_{d+1}(\tilde{\lambda}_1,...,\tilde{\lambda}_l,\tilde{\lambda}_{l+1};n_1,...,n_l,1)=\frac{1}{\prod_{i=1}^{l}\Gamma(n_i)}\frac{\partial^{n_1-1}\cdots\partial^{n_l-1}}{\partial^{n_1-1}\tilde{\lambda}_1\cdots\partial^{n_l-1}\tilde{\lambda}_l}\times\left[\prod_{i=1}^{l+1}\tilde{\lambda}_i^{n_i-1}\phi^{\sigma}_{l+1}(\tilde{\lambda}_1,...,\tilde{\lambda}_{l+1})\right]$$

and this concludes the proof. □

Proof of Theorem 2. The d-dimensional tail integral is

$$U(x_1,...,x_d)=\int_{x1}^{+\infty}\int_{x_2}^{+\infty}\cdots\int_{x_d}^{+\infty}(-1)^{d-1}\nu^{(d-1)}(y_1+\cdots+y_d)dy_d\cdots dy_2dy_1$$

The following change of variable, $z_1=y_1, z_2=y_2+y_1, ..., z_d=y_d+\cdots+y_1$, yields

$$\begin{aligned}U(x_1,...,x_d)&=\int_{x_1+\cdots+x_d}^{+\infty}\int_{x_1+\cdots+x_{d-1}}^{z_d-x_d}\cdots\int_{x_1}^{z_2-x_2}(-1)^{d-1}\nu^{(d-1)}(z_d)dz_1\cdots dz_{d-1}dz_d\\&=\int_{x_1+\cdots+x_d}^{+\infty}(-1)^{d-1}\nu^{(d-1)}(z_d)\int_{x_1+\cdots+x_{d-1}}^{z_d-x_d}\cdots\int_{x_1}^{z_2-x_2}dz_1\cdots dz_{d-1}dz_d\\&=\int_{x_1+\cdots+x_d}^{+\infty}(-1)^{d-1}\nu^{(d-1)}(z_d)\frac{1}{(d-1)!}\left[z_d-(x_1+\cdots+x_d)\right]^{d-1}dz_d\end{aligned}$$

We analyze the above integral through integration by parts and obtain

$$\begin{aligned}U(x_1,...,x_d)=&\frac{1}{(d-1)!}\left[z_d-(x_1+\cdots+x_d)\right]^{d-1}\nu^{(d-2)}(z_d)\Big|_{x_1+\cdots+x_d}^{+\infty}\\&-\int_{x_1+\cdots+x_d}^{+\infty}\frac{1}{(d-2)!}\left[z_d-(x_1+\cdots+x_d)\right]^{d-2}(-1)^{d-1}v^{(d-2)}(z_d)dz_d\end{aligned}$$

Note that (2.12) implies that $\lim_{z_d\to\infty}\left[z_d-(x_1+\cdots+x_d)\right]^{d-1}\nu^{(d-2)}(z_d)=0$, we have

$$U(x_1,...,x_d)=\int_{x_1+\cdots+x_d}^{+\infty}\frac{1}{(d-2)!}\left[z_d-(x_1+\cdots+x_d)\right]^{d-2}(-1)^{d-2}v^{(d-2)}(z_d)dz_d$$

Repeat $(d-2)$ times the same strategy of integration by parts and the implication

of (2.12), we could get

$$U(x_1, ..., x_d) = \int_{x_1+\cdots+x_d}^{+\infty} \nu(z_d)dz_d = U(x_1 + \cdots + x_d)$$

According to Theorem 5.3 in Cont and Tankov (2004), there is only one copula such that

$$U(x_1, ..., x_d) = C(U(x_1), ..., U(x_d))$$

Since $U(x_1, ..., x_d) = U(x_1 + \cdots + x_d)$, it's easy to see that

$$C(y_1, ..., y_d) = U(U^{-1}(y_1) + \cdots + U^{-1}(y_d))$$

□

Proof of Theorem 3. We recall that with $\tilde{\mu}_i$, $i = 1, ..., d$, we denote the i-th σ-stable completely random measure, see Section 2.

$$\tilde{\pi}_k^{n_1,...,n_d}(\boldsymbol{n_1, ..., n_d}, d\boldsymbol{z}) = \frac{\Gamma^d(\theta)}{K \prod_{i=1}^{d} [\tilde{\mu}_i(\mathbb{X})]^{\theta+n_i}} \prod_{j=1}^{k} [\tilde{\mu}_1(dz_j)]^{n_{j,1}} \cdot \cdots \cdot [\tilde{\mu}_d(dz_j)]^{n_{j,d}}$$

for any $k \geq 1$ and $\boldsymbol{n_i} = (n_{1,i}, ..., n_{k,i})$ such that $\sum_{j=1}^{k} n_{j,i} = n_i$ for $i = 1, ..., d$. We will now show that the probability distribution $\mathbb{E}[\tilde{\pi}_k^{n_1,...,n_d}]$ admites a density on $\mathbb{N}^{dk} \times \mathbb{X}^k$ with respect to the product measure $\gamma^{dk} \times \alpha^k$, where γ is the counting measure on the positive integers, and will determine its form. Suppose $C_{\epsilon,x}$ denotes a neighborhood of $x \in \mathbb{X}$ of radius $\epsilon > 0$ and $B_\epsilon = \cup_{j=1}^{k} C_{\epsilon,z_j}$. Then

$$\begin{aligned}\int_{B_\epsilon} \mathbb{E}[\tilde{\pi}_k^{n_1,...,n_d}(\boldsymbol{n_1, ..., n_d}, d\boldsymbol{z})] = \frac{1}{K \prod_{i=1}^{d} (\theta)_{n_i}} \int_{(0,\infty)^d} \lambda_1^{\theta+n_1-1} \cdot \cdots \cdot \lambda_d^{\theta+n_d-1} \\ \times \mathbb{E}\left[e^{-\lambda_1\tilde{\mu}_1(\mathbb{X})-\cdots-\lambda_d\tilde{\mu}_d(\mathbb{X})} \prod_{j=1}^{k} \prod_{i=1}^{d} \left[\tilde{\mu}_i(C_{\epsilon,z_j})\right]^{n_{j,i}} \right] d\boldsymbol{\lambda}\end{aligned}$$

Define $\mathbb{X}_\epsilon$ to be the whole space $\mathbb{X}$ with the neighbourhoods C_{ϵ,z_r} deleted for all $j = 1, ..., k$. By virtue of the independence of the increments of the CRMs $\tilde{\mu}_1$ and

$\tilde{\mu}_2$, the expression above reduces to

$$\frac{1}{K\prod_{i=1}^{d}(\theta)_{n_i}}\int_{(0,\infty)^d}\lambda_1^{\theta+n_1-1}\cdots\lambda_d^{\theta+n_d-1}\mathbb{E}\left[e^{-\lambda_1\tilde{\mu}_1(\mathbb{X}_\epsilon)-\cdots-\lambda_d\tilde{\mu}_d(\mathbb{X}_\epsilon)}\right]\times\prod_{j=1}^{k}M_{j,\epsilon}(\boldsymbol{\lambda})d\boldsymbol{\lambda}$$

where, by virtue of Lemma 1,

$$\begin{aligned}M_{j,\epsilon}(\boldsymbol{\lambda}) :&= \mathbb{E}\left[e^{-\lambda_1\tilde{\mu}_1(\mathbb{X}_\epsilon)-\cdots-\lambda_d\tilde{\mu}_d(\mathbb{X}_\epsilon)}\prod_{i=1}^{d}\left[\tilde{\mu}_i(C_{\epsilon,z_j})\right]^{n_{j,i}}\right]\\&= \alpha(C_{\epsilon,z_j})e^{-\alpha(C_{\epsilon,z_j})\psi_{\rho,d}(\boldsymbol{\lambda})}g_\rho(n_{j,1},...,n_{j,d};\boldsymbol{\lambda})+o(\alpha(C_{\epsilon,z_j}))\end{aligned}$$

This shows that $\mathbb{E}[\tilde{\pi}^k]$ admits a density with respect to $\gamma^{dk}\times\alpha^k$ and it is given by

$$\frac{1}{K\prod_{i=1}^{d}(\theta)_{n_i}}\int_{(0,\infty)^d}\lambda_1^{\theta+n_1-1}\cdots\lambda_d^{\theta+n_d-1}e^{-\psi_{\rho,d}(\boldsymbol{\lambda})}\times\prod_{j=1}^{k}g_\rho(n_{j,1},...,n_{j,d};\boldsymbol{\lambda})d\boldsymbol{\lambda}$$

□

Appendix B

Proofs of Chapter 3

We derive the analytical form for the formula (3.6). Let K be the number of non-zero clusters, i.e. $K = |\Pi_1(C_{-t,\nu}, j)|$, then

$$
\begin{aligned}
&P\{\mathbf{Y}_\nu(T)|C_{t,\nu} = j, C_{-t,\nu}, \gamma_\nu(T), W_\nu(T)\} \\
&\qquad = \Big\{ \prod_{l\in\Pi_0(C_{-t,\nu},j)} p(Y_{l,\nu}|0) \Big\} \prod_{k=1}^{K} \int \prod_{l\in\Pi_1(C_{-t,\nu},j)_k} p(Y_{l,\nu}|\mu^*_{j,\nu}) G_0(d\mu^*_{j,\nu}) \\
&\qquad \propto \exp\left\{ -\frac{\mu_0^2}{2\sigma_0^2} K + \frac{1}{2} \sum_{k=1}^{K} \frac{(\frac{\mu_0}{\sigma_0^2} + \sum_{l\in\Pi_1(C_{-t,\nu},j)_k} \frac{y_l}{\tau^2})^2}{\frac{1}{\sigma_0^2} + \frac{|\Pi_1(C_{-t,\nu},j)_k|}{\tau^2}} \right\} \prod_{k=1}^{K} \frac{1}{\sqrt{\frac{|\Pi_1(C_{-t,\nu},j)_k|}{\tau^2} + 1}}
\end{aligned}
$$

Acknowledgments

I would like to thank my supervisors Fabrizio Leisen and Juan Miguel Marín for their invaluable support and guidance: I feel blessed to have the opportunity to work with them.
I also would like to express my gratitude to Michele Guindani, who provided me a great chance to work with him in MD Anderson Cancer Center, to Alberto Cassese and Marina Vannucci in Rice University for their precious advice.
My gratitude goes to the Statistics Department in University of Kent for hosting me as a visiting scholar in February-March 2014, and goes to my supervisor Fabrizio again for making all this happen.
I am grateful to the organizers of 9th BNP conference and 4th MCMSki conference for the travel support they provided.
I am glad I spent the last five years working in a friendly environment such as the department of statistics of the University Carlos III de Madrid.
Finally, I thank my wonderful family and my girlfriend Xiaoling Mei for their unconditional support.

Bibliography

Airoldi, E. M., Costa, T., Bassetti, F., Leisen, F., and Guindani, M. (2014). Generalized species sampling priors with latent Beta reinforcements. *Journal of the American Statistical Association*, **109(508)**, 1466-1480.

Antoniak, C.E. (1974), 'Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems', *Ann. Statist*, **2, no. 6**, 1152-1174

Ausin, M.C., Galeano, P., and Ghosh, P. (2014), 'A semiparametric Bayesian approach to the analysis of financial time series with applications to Value at Risk estimation', *European Journal of Operational Research*, **232**, 350-358.

Bassetti, F., Casarin, R., and Leisen, F., (2014), 'Beta-Product dependent Pitman-Yor Processes for Bayesian inference', *Journal of Econometrics*, **180**, 49-72.

Bassetti, F., Crimaldi, I., and Leisen, F. (2010). Conditionally identically distributed species sampling sequences. *Advances in applied probability*, **42(2)**, 433-459.

Beaumont, M., Zhang, W. and Balding, D. (2002). Approximate Bayesian computation in population genetics. *Genetics*, **162(4)**, 2025-2035.

Besag, J. (1977). Efficiency of pseudolikelihood estimation for simple Gaussian fields. *Biometrika*, **64(3)**, 616–618.

Besag, J., York, J. and Mollie, A. (1991). Bayesian image restoration, with two applications in spatial statistics (with discussion). *Ann. Inst. Stat. Math*, **43(1)**, 1–59.

Blackwell, D., and MacQueen, J. B. (1973). 'Ferguson distributions via Pólya urn schemes'. *The annals of statistics*, 353-355.

Bollerslev, T. (1986). Generalized Autorregressive Conditional Heteroskedasticity. *J. Econometrics, 31(3)*, 307–327.

Brouste, A., Fukasawa, M., Hino, H., Iacus, S.M., Kamatani, K., Koike, Y., Masuda, H., Nomura, R., Ogihara, T., Shimuzu, Y., Uchida, M. and Yoshida, N. (2014). The YUIMA Project: A Computational Framework for Simulation and Inference of Stochastic Differential Equations. *J. Stat. Software*, **57(4)**, 1–51.

Buck, C.E., Cavanagh, W.G. and Litton, C.D. (1988). The spatial analysis of site phosphate data. In: Rhatz, S.P.Q. (ed.) Computer Applications and Quantitive Methods in Archeology. British Archaeological Reports, International Series, **446**, BAR, Oxford.

Cabras, S., Nueda, M.E.C., Ruli, E., Approximate Bayesian Computation by Modelling Summary Statistics in a Quasi-likelihood Framework, *Bayesian Analysis*, **10(2)**, 411-439.

Constantines G.M., Savits T.H. (1996). 'A multivariate version of the Faa di Bruno formula'. *Trans. Amer. Math. Soc.*, **348**, 503–520.

Cont, R., and Tankov, P., (2004), *Financial modelling with jump processes*, Boca Raton, FL: Chapman & Hall/CRC.

Cornuet, J.M., Pudlo, P., Veyssier, J., Dehne-Garcia. A., Gautier, M., Leblois, R., Marin, J.M. and Estoup, A. (2014). DIYABC v2.0: a software to make Approximate Bayesian Computation inferences about population history using Single Nucleotide Polymorphism, DNA sequence and microsatellite data. *Bioinformatics*, **30(8)**, 1187–1189.

Davison, A.C., Hinkley, D.V. and Worton, B.J. (1992). Bootstrap likelihoods. *Biometrika*, **79(1)**, 113–130.

Davison, A.C. and Hinkley, D.V. (1997). Bootstrap methods and their application. Cambridge University Press.

Dean, T., Singh, S., Jasra, A. and Peters, G. (2014). Estimation of HMMs with Intractable Likelihoods. *Scand. J. Statist.* **41(4)**, 970–987.

Drovandi, C. and Pettitt, A. (2010). Estimation of parameters for macroparasite population evolution using approximate Bayesian computation. *Biometrics*, **67(1)**, 225–233.

Daley, D.J., and Vere-Jones, D., (2003), *An introduction to the theory of point processes. Vol. 1.*, New York: Springer.

De Iorio, M., Müller, P., Rosner, G.L., and MacEachern, S.N., (2004), 'An ANOVA model for dependent random measures', *J. Amer. Statist. Assoc.*, **99**, 205–215.

Dean, T., Singh, S., Jasra, A. and Peters, G. (2014). Estimation of HMMs with Intractable Likelihoods. *Scand. J. Statist.* (to appear).

Del Moral, P., Doucet, A. and Jasra, A. (2012). An adaptive sequential Monte Carlo method for approximate Bayesian computation. *Stat. Comput.* **22**, 1009-1020.

Drovandi, C. and Pettitt, A. (2010). Estimation of parameters for macroparasite population evolution using approximate Bayesian computation. *Biometrics*, **67(1)**, 225-233.

Escobar, D. (1994), 'Estimating normal means with a Dirichlet process prior', *J. Amer. Statist. Assoc.*, **89**, 268-277.

Escobar, M.D., and West, M., (1995), 'Bayesian density estimation and inference using mixtures', *J. Amer. Statist. Assoc.*, **90**, 577-588.

Favaro, S., Lijoi, A., Mena, R.H., and Pruenster, I., (2009), 'Bayesian nonparametric inference for species variety with a two parameter Poisson-Dirichlet process prior', *Journal of the Royal Statistical Society Series B*, **71**, 993-1008.

Favaro, S., Lijoi, A., and Pruenster, I., (2012), 'A new estimator of the discovery probability', *Biometrics*, **68**, pp. 1188-1196.

Fearnhead, P. and Prangle, D. (2012). Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. *J. Royal Statistical Society (B)*, **74(3)**, 419-474.

Ferguson, T.S., (1973), 'A Bayesian analysis of some nonparametric problems', *Ann. Statist.*, **1**, 209–230.

Griffin, J.E., (2011), 'Inference in infinite superpositions of Ornstein-Uhlenbeck processes using Bayesian non-parametrics methods', *Journal of Financial Econometrics*, **1**, 1–31.

Griffin J E, Leisen F (2014), Compound random measures and their use in Bayesian nonparametrics[J]. *arXiv preprint* arXiv:1410.0611.

Griffin, J.E., and Steel, M.F.J., (2006), 'Order-based dependent Dirichlet processes', *Journal of the American Statistical Association*, **101**, 179–194.

Griffin, J.E., and Steel, M.F.J., (2011), 'Stick Breaking Autoregressive Processes', *Journal of Econometrics*, **162**, 383–396.

Griffin, J.E., and Walker, S. G., (2012), 'Posterior Simulation of Normalized Random Measures Mixtures', *Journal of Computational and Graphical Statistics* **20**, 241–259.

Griffin, J. E., Kolossiatis, M., and Steel, M. F. (2013), 'Comparing distributions by using dependent normalized random-measure mixtures', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **75(3)**, 499-529.

Grimmett, G. (2010). Probability on Graphs Random Processes on Graphs and Lattices. Cambridge University Press.

Hatjispyrosa, S.J., Nicoleris, T.N., and Walker, S.G., (2011), 'Dependent mixtures of Dirichlet processes', *Computational Statistics and Data Analysis*, **55**, 2011–2025.

Huettel, S.A., Song A.W., McCarthy G. (2004), Functional Magnetic Resonance Imaging, vol. 1. Sunderland, MA: Sinauer Associates.

Ishwaran, H., James, L.F., (2001), 'Gibbs sampling methods for stick-breaking priors', *J. Amer. Stat. Assoc.*, **96**, 161–173.

Kalli, M, Griffin, J.E., and Walker, S.G., (2011), 'Slice Sampling Mixture Models', *Statistics and Computing*, **21**, 93–105.

J.F.C. Kingman (1975). Kingman, J. F. C. (1975). Random discrete distributions (with discussion). *J. Roy. Statist. Soc. Ser. B* **37**, 1–22.

Kingman, J. (1967), 'Completely random measures', *Pacific Journal of Mathematics*, **21(1)**, 59-78.

Kingman, J.F.C., (1993), *Poisson processes*, Oxford: Oxford University Press.

Kolossiatis, M., Griffin, J.E., and Steel, M.F.J., (2013), 'On Bayesian nonparametric modelling of two correlated distributions', *Statistics and Computing*, **23**, 1–15.

Lahiri, S.N. (2003). Resampling Methods for Dependent data. Springer–Verlag, New York.

Leisen, F., and Lijoi, A., (2011), 'Vectors of Poisson-Dirichlet processes', *J. Multivariate Anal.*, **102**, 482–495.

Leisen, F., Lijoi, A., and Spano, D., (2013), 'A Vector of Dirichlet processes', *Electronic Journal of Statistics*, **7**, 62–90.

Lijoi, A. and Nipoti, B. (2014), 'A class of hazard rate mixtures for combining survival data from different experiments', *Journal of the American Statistical Association*, **109**, 802-814.

Lijoi, A., Nipoti, B., and Prünster, I. (2014a), 'Bayesian inference with dependent normalized completely random measures', *Bernoulli*, **20**, 1260-1291.

Lijoi, A., Nipoti, B., and Prünster, I. (2014b), 'Dependent mixture models: clustering and borrowing information', *Computational Statistics and Data Analysis*, **71**, 417-433.

Lo, A.Y. (1984). On a class of Bayesian nonparametric estimates. I. Density estimates. *Ann. Statist.* **12**, 351–357.

MacEachern, S.N., (1999), 'Dependent nonparametric processes'. *ASA Proceedings of the Section on Bayesian Statistical Science*, 50–55.

Marin, J.M. and Robert, C.P. (2014). Bayesian Essentials with R. Springer–Verlag, New York.

Marin, J.M., Pudlo, P., Robert, C.P., Ryder R. J. (2012). Approximate Bayesian computational methods. *Statistics and Computing*, **22(6)**, 1167–1180.

McGrory, C.A., Titterington, D.M., Reeves, R. and Pettitt, A.N. (2009). Variational Bayes for estimating the parameters of a hidden Potts model. *Stat. Comput.*, **19(3)**, 329–340.

McKinley, T., Cook, A. and Deardon, R. (2009). Inference in epidemic models without likelihoods. *Int. J. Biostat.*, **5(1)**, 1–40.

Marin, J.M., Pudlo, P., Robert, C.P., Ryder R. J. (2012). Approximate Bayesian computational methods. *Statistics and Computing*, **22(6)**, 1167–1180.

McKinley, T., Cook, A. and Deardon, R.(2009). Inference in epidemic models without likelihoods. *Int. J. Biostat.*, **5(1)**, 24.

Neal, R.M., (2000), 'Markov chain sampling methods for Dirichlet process mixture models', *Journal of Computational and Graphical Statistics*, **9**, 249-265.

Mengersen, K.L., Pudlo, P. and Robert, C.P. (2013). Bayesian computation via empirical likelihood. *Proceedings of the National Academy of Sciences*, **110(4)**, 1321-1326.

Olkin, I. and Liu, R. (2003). A bivariate beta distribution. *Stat. Prob. Letters*, **62(4)**, 407–412.

Owen, A.B. Empirical likelihood. CRC press, 2010.

Picchini, U. (2014). Inference for SDE Models via Approximate Bayesian Computation. *J. of Comp. and Graph. Stat.*, **23(4)**, 1080–1100.

J. Pitman (1995). Exchangeable and partially exchangeable random partitions *Probab. Theory Related Fields.* **102**, 145–158.

Pitman, J., (2006), *Combinatorial Stochastic Processes*, Berlin: Springer.

Pitman, J. and Yor, M. (1997), 'The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator', *Ann. Probab.*, **25**, 855–900.

Pritchard, J., Seielstad, M., Perez-Lezaun, A. and Feldman, M. (1999). Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular Biology and Evolution*, **16(12)**, 1791–1798

Sethuraman J (1991). 'A constructive definition of Dirichlet priors'. *Florida State Univ Tallahassee Dept of Statistics.*

Sisson, S.A., Fan, Y. and Tanaka, M. (2007). Sequential Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci.*, **104(6)**, 1760-1765.

Wilson, I.J. and Balding, D.J. (1998). Genealogical inference from microsatellite data. *Genetics*, **150(1)**, 499–510.

Zhu, W., and Leisen, F. (2014). A multivariate extension of a vector of two-parameter Poisson-Dirichlet processes. *Journal of Nonparametric Statistics*, **27-1**, 89-105.

Zhu, J. and Morgan, G.D. (2004). Comparison of Spatial Variables over Subregions Using a Block Bootstrap. *J. of Agricultural, Biological, and Environmental Statistics.*, **9(1)**, 91–104.